DISTRIBUTIONALLY ROBUST STOCHASTIC OPTIMIZATION WITH APPLICATIONS IN STATISTICAL LEARNING

A Dissertation
Presented to
The Academic Faculty

By

Rui Gao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial & Systems Engineering

Georgia Institute of Technology

May 2018

DISTRIBUTIONALLY ROBUST STOCHASTIC OPTIMIZATION WITH APPLICATIONS IN STATISTICAL LEARNING

Approved by:

Dr. Anton J. Kleywegt, Advisor
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Shabbir Ahmed
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Jim Dai
School of Operations Research and Information Engineering
Cornell University

Dr. Alexander Shapiro
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Melvyn Sim
NUS Business School
National University of Singapore

Date Approved: April 5, 2018

When one admits that nothing is certain one must, I think, also admit that some things are much more nearly certain than others.

— Bertrand Russell

To my parents, grandmother and the love of my life

ACKNOWLEDGEMENTS

My foremost and deepest gratitudes goes to my advisor, Anton Kleywegt. I have been very fortunate to work with him on exciting problems. It is hard to imagine this thesis without his generous support, invaluable guidance and selfless dedication. I cannot overstate my appreciation for his constant availability, perpetual patience, enormous encouragement, and the flexible research atmosphere he created. Discussing research problems with him (day and night) is one of the most rewarding and enjoyable things during my PhD. I owe him a debt of gratitude larger than I can express here.

I am very much indebted to all my committee members, Shabbir Ahmed, Jim Dai, Alex Shapiro, and Melvyn Sim, for their valuable comments on the thesis and generous help during my job search. Shabbir has always been a tremendous source of support and keeps providing me with valuable feedbacks on my research. Thanks to Jim, I was fortunate to have the opportunity to work on the airline revenue management project, and I benefited a lot from his insightful career advice and video lectures. My interest in stochastic optimization stems from reading Alex's book when I was an undergraduate student, since then his knowledge and wisdom have been constantly shaped my point of view. I am very thankful to Melvyn for offering valuable advice on my future direction, sharing his research philosophy, and inviting me to NUS.

Besides my thesis committee, I would like to thank many outstanding faculty members at Georgia Tech. In particular, I would like to give my heartfelt and special thanks to David Goldberg for his tremendous advice and constructive feedbacks. Conversations with him have always inspired me to be a better researcher. I benefited a lot from unforgettable lectures delivered by Santanu Dey, Ton Dieker, Bob Foley, Wilfrid Gangbo, George Nemhauser, Arkadi Nemirovski, Craig Tovey and Jeff Wu, as well as those from Ronald Johnson, Judith Norback and Damon Williams who help improve my teaching and communication skills. I am indebted to Eva Lee, without whom I would never be able to start

my PhD at ISyE. Futhermore, I am very grateful to Hayriye Ayhan, Alan Erera, Siva Theja Maguluri, Edwin Romeijn, Julie Swann, Andy Sun, Alejandro Toriello, He Wang, Yao Xie, Huan Xu, and Tuo Zhao for their generosity, support and help on my studies, research and/or job search.

In addition, I would like to thank the entire Stochastic/Robust Optimization Community, for welcoming me with such open arms. I look forward to the professional contact with you all in the years to come. Although there are too many names to mention here, I would like to thank my coauthors, who I learnt a lot from, Xi Chen, Feng Qiu, Cheng Wang, and Linwei Xin.

My PhD life is certainly less colorful without my fellow students and friends at Georgia Tech. I appreciate all their accompany, support, and assistant. Special thanks go to Yang Cao, Yufeng Cao, Junzhuo Chen, Yilun Chen, Weijun Ding, Zhihao Ding, Xiaolei Fang, Junxuan Li, Yuan Li, Zihao Li, Yifan Liu, Simon Mak, Xinyu Min, Jan Vlachy, Wenjia Wang, Xin Wang, Xinchang Wang, Yichen Wang, Weijun Xie, Fan Ye, Xiaowei Yue, Can Zhang, Yi Zhang, Yi Zhou, Helin Zhu, Yuanji Zhu, and Jikai Zou.

None of this would have been possible without the continuous and unconditional love and support of my dearest parents and grandmother. I owe them more than I could ever express here. I would also like to thank the love of my life, Lu Lei, who has always been, and continues to be, the primary source of peace and happiness in my life. Through their many sacrifices, I have been able to proceed during this long journey. To them I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In this thesis, we study distributionally robust stochastic optimization (DRSO), a recent emerging framework for solving decision-making under uncertainty. In this framework, instead of assuming that there is a known underlying probability distribution that drives the uncertain behavior of stochastic systems, one seeks solutions that perform well for a family of distributions, so as to hedge against the distributional uncertainty in the future. This thesis focuses on the design of tractable models for DRSO. We develop novel formulations and insights for fundamental problems, and discover connections between different areas in optimization, statistics and learning.

We first address the key question on how to construct a good family of distributions to hedge against. We point out that such family should be chosen to be appropriate for the application at hand, and that some of the choices that have been popular until recently are, for many applications, not good choices. We consider distributions that are within a chosen Wasserstein distance from a nominal distribution, for example an empirical distribution resulting from available data. We demonstrate that the resulting distributions hedged against are more reasonable than those resulting from other popular choices of sets. Moreover, the problem of determining the worst-case expectation over the resulting family of distributions has desirable tractability properties. We derive a dual reformulation of the Wasserstein DRSO problem in a very general setting, by constructing (approximate) worst-case distributions explicitly via the first-order optimality conditions of the dual problem. By construction, the worst-case distributions have a concise structure and a clear interpretation.

Next, we establish a connection between Wasserstein DRSO and regularization in statistical learning. More precisely, we identify a broad class of loss functions, for which the Wasserstein DRSO is asymptotically equivalent to a regularization problem with a gradient-norm penalty. Such relation provides new interpretations for problems involv-

ing regularization, including a great number of statistical learning problems and discrete choice models (e.g. multinomial logit). The connection also suggests a principled way to regularize high-dimensional non-convex learning problems, which is demonstrated through the training of Wasserstein generative adversarial networks in deep learning.

In the final part of the thesis, we consider robust decision-making when the data availability from marginal distributions is different than that from the joint distribution. This occurs, for example, when the data streams of different random variables are collected with different frequencies. We propose a distributionally robust approach which hedges against a family of joint distributions with fixed marginals and a dependence structure similar to that of a nominal joint distribution, such as an empirical distribution or the independent product distribution. Similarity of the dependence structure is measured through the Wasserstein distance between the copula of the joint distribution and the copula of the nominal distribution. We show that our choice of distance can be used as a new measure of dependence among random variables. Tractability of our new formulation is obtained by a novel constructive proof of strong duality, combining ideas from variational analysis and the theory of multi-marginal optimal transport.

CHAPTER 1
INTRODUCTION

Decision-making under uncertainty problems occur in numerous fields of science, engineering and management. This stimulates theoretical and practical interest from diverse research communities in operations research, statistics, and machine learning. Throughout the years, several solution approaches have been proposed to formulate, analyze and solve these problems, including stochastic optimization, robust optimization, dynamic programming, etc. Traditionally, these optimization models describes the uncertainty via probability distributions which we assume can be estimated accurately from the observed data. However, such assumption are being challenged especially in the era of Big Data. Indeed, the trend of Big Data is towards more observations but even more so, to a greater number of uncertain variables. Such problems are often called high-dimensional problems in statistics, and can be met in e-commerce (tracking, loyalty programs, etc.), finance and economics, medicine (medical images, biotech data), to name a few. The challenge for these problems is that, the nominal probability distribution constructed from the observed data (e.g. empirical distribution or some fitted distribution) may be not representative of the underlying true distribution. Strategies solely based on the nominal distribution can lead to poor performance when implemented, as the distributional uncertainty may be amplified through the optimization process. Furthermore, the future uncertainty can sometimes be adversarially different from the observed data, for instance, when an attacker intentionally design inputs (e.g. email spam) to cause machine learning models (e.g. spam filter) to make mistakes. Therefore, a fundamental problem in decision-making under uncertainty as well as in statistical learning is:

> *How can we find solutions that perform well not only on the observed data, but also generalize to new and previously unseen data?*

In this thesis, we aim to answer this fundamental problem by advancing methods in the emerging field of *distributionally robust stochastic optimization* (DRSO). We focus on designing tractable models that hedges against different types of distributional uncertainty. Leveraging tools from probability, optimal transport and variational analysis, we show that our proposed models (based on *Wasserstein distance*) enjoy nice tractability and are able to handle uncertainty due to *high dimensionality*, *data perturbation*, and *data availability*. In addition, we establish connections between DRSO and *regularization*, a classical approach in statistical learning to designing models and algorithms with good generalization ability.

1.1 Distributionally Robust Stochastic Optimization

In decision making problems under uncertainty, a decision maker wants to choose a decision β from a feasible region $\mathcal{D}$. The objective function $\ell : \mathcal{D} \times \Xi \to \mathbb{R}$ depends on a quantity $\boldsymbol{\xi} \in \Xi$ whose value is not known to the decision maker at the time that the decision has to be made. In some settings it is reasonable to assume that $\boldsymbol{\xi}$ is a random element with distribution $\boldsymbol{\mu}$ supported on Ξ, for example, if multiple realizations of $\boldsymbol{\xi}$ will be encountered. In such settings, the decision making problems can be formulated as a *stochastic optimization* problem:

$$\inf_{\beta \in \mathcal{D}} \ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta, \boldsymbol{\xi})].$$

We refer to [1] for a thorough study of stochastic optimization.

As mentioned above, one major criticism of the stochastic optimization formulation for practical applications is the requirement that the underlying distribution $\boldsymbol{\mu}$ be known to the decision maker. Even if multiple realizations of $\boldsymbol{\xi}$ are observed, $\boldsymbol{\mu}$ still may not be known exactly, while use of a distribution different from $\boldsymbol{\mu}$ may sometimes result in bad decisions. Another major criticism is that in many applications there are not multiple realizations of $\boldsymbol{\xi}$ that will be encountered, for example in adversarial attacks or problems involving events that may either happen once or not happen at all, and thus the notion of a "true" underlying

distribution does not apply. These criticisms motivate the notion of *distributionally robust stochastic optimization* (DRSO), that does not rely on the notion of a known true underlying distribution. One chooses a set $\mathfrak{M}$ of probability distributions to hedge against, and then finds a decision that provides the best hedge against the set $\mathfrak{M}$ of distributions by solving the following minmax problem:

$$\inf_{\beta \in \mathcal{D}} \sup_{\boldsymbol{\mu} \in \mathfrak{M}} \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta, \boldsymbol{\xi})]. \tag{DRSO}$$

Such a minmax approach has its roots in Von Neumann's game theory and has been used in many fields such as inventory management [2, 3], statistical decision analysis [4], as well as stochastic optimization [5, 6, 7]. Recently it regained attention in the operations research and machine learning, and sometimes is called data-driven stochastic optimization or ambiguous stochastic optimization.

A central question is: how to choose a good set of distributions $\mathfrak{M}$ to hedge against? A good choice of $\mathfrak{M}$ should take into account the properties of the practical application as well as the tractability of problem (DRSO). Two typical ways of constructing $\mathfrak{M}$ are *moment-based* and *distance-based*. The moment-based approach considers distributions whose moments (such as mean and covariance) satisfy certain conditions [2, 8, 9, 10]. It has been shown that in many cases the resulting DRSO problem can be formulated as a conic quadratic or semi-definite program. However, the moment-based approach is based on the curious assumption that certain conditions on the moments are known exactly but that nothing else about the relevant distribution is known. More often in applications, either one has data from repeated observations of the quantity $\boldsymbol{\xi}$, or one has no data, and in both cases the moment conditions do not describe exactly what is known about $\boldsymbol{\xi}$. In addition, the resulting worst-case distributions sometimes yield overly conservative decisions [11, 12]. For example, [11] shows that for the newsvendor problem, by hedging against all the distributions with fixed mean and variance, Scarf's moment approach yields a two-point

worst-case distribution, and the resulting decision does not perform well under other more likely scenarios. The above issues can be partially resolved by considering the generalized moment approach [13, 14], but is beyond the scope of our discussion. The distance-based approach considers distributions that are close, in the sense of a chosen statistical distance, to a *nominal distribution* $\boldsymbol{\nu}$, such as an empirical distribution or a fitted Gaussian distribution [15, 16]. Popular choices of the statistical distance are ϕ-divergences [17, 18], which include Kullback-Leibler divergence [19], Burg entropy [11], and Total Variation distance [20] as special cases, Prokhorov metric [21], and Wasserstein distance [22, 23, 24, 25]. For Prokhorov metric, the resulting DRSO problem is tractable in rare cases [21]. For divergence measures, we postpone the discussion of their potential issues in Chapter 2.

1.2 Problem Formulation and Literature Review

1.2.1 Distributionally Robust Stochastic Optimization with Wasserstein distance

We mainly focus on a family of distribution based on the *Wasserstein distance*. Specifically, consider any underlying metric d on Ξ which measures the closeness of any two points in Ξ. Let $p \geq 1$, and let $\mathcal{P}(\Xi)$ denote the set of Borel probability measures on Ξ. The Wasserstein distance of order p between two distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\Xi)$ is defined as

$$\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \;:=\; \min_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)} \Big\{ \mathbb{E}^{1/p}_{(\boldsymbol{\xi},\boldsymbol{\zeta}) \sim \boldsymbol{\gamma}} [\mathsf{d}^p(\boldsymbol{\xi}, \boldsymbol{\zeta})] \;:\; \boldsymbol{\gamma} \text{ has marginal distributions } \boldsymbol{\mu}, \boldsymbol{\nu} \Big\}.$$

More detailed explanation and discussion on Wasserstein distance will be presented in Section 2.3. Given a nominal distribution $\boldsymbol{\nu}$ and a radius $\theta > 0$, we are interested in solving

$$\min_{\beta \in \mathcal{D}} \sup_{\boldsymbol{\mu} \in \mathcal{P}(\Xi)} \big\{ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta, \boldsymbol{\xi})] : \; \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \theta \big\}. \qquad \text{(Wasserstein-DRSO)}$$

Wasserstein distance and the related field of *optimal transport*, which is a generaliza-

tion of the transportation problem, have been studied in depth. In 1942, together with the linear programming problem [26], Leonid Kantorovich [27] tackled Monge's problem originally brought up in the study of optimal transport. In the stochastic optimization literature, Wasserstein distance has been used for single stage stochastic optimization [22, 23], and for multistage stochastic optimization [28]. The challenge for solving (Wasserstein-DRSO) is that, the inner maximization involves a supremum over possibly an infinite dimensional space of distributions. To tackle this problem, existing works focus on the setup when $\boldsymbol{\nu}$ is the empirical distribution on a finite-dimensional space. Particularly, [22] transformed the inner maximization problem of (Wasserstein-DRSO) into a finite-dimensional nonconvex program, by using the fact that if $\boldsymbol{\nu}$ is supported on at most n points, then there are extreme distributions of the Wasserstein ball that are supported on at most $n + 3$ points. Recently, using duality theory of conic linear programming [29], [24] and [25] showed that under certain conditions, the inner maximization problem of (Wasserstein-DRSO) is actually equivalent to a finite-dimensional convex problem.

1.2.2 Distributional Robustness and Regularization

In statistical learning, *regularization* is a typical approach to improve the generalization ability of learning models. Given the empirical distribution $\boldsymbol{\nu}_n$, one consider a regularized stochastic optimization problem

$$\min_{\beta \in \mathcal{D}} \ \mathbb{E}_{\boldsymbol{\nu}_n}[\ell(\beta, \boldsymbol{\xi})] + \theta \cdot J(\beta), \qquad \text{(Regularization)}$$

where θ is the tunning parameter and J is the regularization penalty function. This formulation not only covers commonly seen norm-penalty regularization methods, such as ℓ_1-regularization [30] and Tikhonov regularization [31], but also is (approximately) equivalent to other regularization methods, including adding noise [32], dropout [33, 34], and adversarial training [35]. We refer to Chapter 7 of [36] for a survey on regularization meth-

ods in machine learning.

It is natural to investigate the connection between the regularization and Wasserstein DRSO problem with centered at the empirical distribution $\boldsymbol{\nu}_n$ (i.e., $\boldsymbol{\nu} = \boldsymbol{\nu}_n$ in the problem (Wasserstein-DRSO)). It has been shown that norm penalty regularization has a *robust-optimization* interpretation in some special cases, including linear/matrix regression [37, 38], and support vector machine [39]. Note that in Chapter 2 we show that the problem (Wasserstein-DRSO) can be approximated by a robust optimization problem. In view of this close relationship between (Wasserstein-DRSO) and robust optimization, it is conceivable that in the above-mentioned special cases, (Wasserstein-DRSO) may also be closely related to norm penalty regularization. Indeed, equivalence between (Wasserstein-DRSO) and regularization has been established in [40] for piecewise-linear convex loss, in [41] for logistic regression, and in [42] for linear regression and support vector machines. In a recent work, [43] studies the equivalence between regularization and DRSO with 1-Wasserstein distance ($p = 1$) for linear function class and its kernelization. We finally remark that besides the Wasserstein DRSO, the equivalence between regularization and DRSO with other distances has also been studied. For example, [44] and [45] have pointed out that DRSO with ϕ-divergence is first-order equivalent to variance regularization.

1.2.3 Distributionally Robust Stochastic Optimization with Known Marginal Distributions

In many applications, the *data availability* from marginal distributions is different than that from the joint distribution. This occurs when the data streams of different random variables are collected with different frequencies, the decision maker may have more data on the marginal distributions than on the joint distribution. Consider the example in [46], in which the decision maker wants to measure the dependence between the lengths of delay of two nonstop flights A and B from Los Angeles to Sydney. One flight operates daily, while the other operates on Mondays, Wednesdays, and Saturdays. Thus, we have joint data on

the lengths of delay of the two flights on the days of week when they both operate, and on the remaining days we have additional data on the length of delay of the flight that operates daily. Now the questions is: how can we find robust solutions when the joint distribution is unknown while the marginal distributions can be estimated rather accurately?

Copula theory [47, 48] provides a unified way to model the multivariate dependence that is applicable to the above data availability regime. A *copula* is a multivariate distribution with all univariate marginals being uniformly distributed on $[0,1]^K$. The seminal Sklar's theorem [49] states that, for every multivariate joint distribution function $F^{\boldsymbol{\mu}}$ with marginal distributions $\{F_k\}_{k=1}^K$, there exists a probability distribution function $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ on $[0,1]^K$, such that

$$F^{\boldsymbol{\mu}}(\xi_1,\ldots,\xi_k) \;=\; \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(F_1(\xi_1),\ldots,F_k(\xi_k)),\;\; \forall\, \xi \in \Xi. \tag{1.1}$$

Such $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ is unique if the marginals are continuous. Conversely, any copula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and marginal distributions $\{F_k\}_k$ together define a K-dimensional joint distribution through (1.1). This result is phenomenal since it suggests that the analysis of the dependence structure of a multivariate joint distribution can be separated from knowledge of the marginal distributions. For a detailed illustration on constructing copula, we refer to Section 4.3.

Using copula theory, the uncertainty of the joint distribution all boils down to uncertainty of the copula, provided that the marginal distributions are known. Then a classical approach to tackling the above robust decision-making problem with know marginals is to formulate a minimax problem which hedges against all probability distributions $\mathcal{P}(\Xi)$ on Ξ with the given marginals:

$$\min_{\beta\in\mathcal{D}} \sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta,\boldsymbol{\xi})] :\; \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}} \right\}, \tag{1.2}$$

where the marginal distribution functions $F_1,\ldots,F_k$ are given; and $\mathfrak{C}$ is the set of all copulas on $[0,1]^K$. Such an approach can be traced back at least to Hoeffding [50] and Fréchet [51], who considered the extremes and bounds of (1.2). Since then, this approach has been

extensively studied and applied to many operations management problems [52, 53, 54]. We refer to [55] and [56] for a thorough study on this topic.

Formulation (1.2) is often very conservative for many interesting applications (see, e.g. Example 4.1), because it only uses the information of marginal distributions. To overcome its over-conservativeness and make a better use of the potentially available joint data, it is natural to restrict $\boldsymbol{\mathcal{C}}$ to a smaller set. Indeed, using the idea from distributionally robust stochastic optimization, recent research considers balls of copulas that are close to some nominal copula $\boldsymbol{\mathcal{C}}^0$ in the sense of Kullback-Leibler (KL) divergence[1] [58, 59, 60, 61]:

$$\min_{\beta\in\mathcal{D}} \sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}} \left\{\mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta,\boldsymbol{\xi})] :\ \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}},\ KL(\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{C}}^0)\le\theta\right\}, \tag{1.3}$$

possibly with some additional constraints. As will be shown in Section 4.2, such choice may have some undesirable properties for data-driven problems.

Motivated by the results in Chapter 2, we consider all distributions whose associated copula is close to some nominal copula $\boldsymbol{\mathcal{C}}^0$ in the Wasserstein distance. More specifically, when there is available joint data, we set $\boldsymbol{\mathcal{C}}^0$ to be the empirical copula, and when there is no joint data, we set $\boldsymbol{\mathcal{C}}^0$ to be the independent copula. Let $\mathcal{W}_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}},\boldsymbol{\mathcal{C}}^0)$ denote the p-Wasserstein distance between $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and $\boldsymbol{\mathcal{C}}^0$. Let $\theta>0$, we propose the following formulation

$$\min_{\beta\in\mathcal{D}} \sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}} \left\{\mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta,\boldsymbol{\xi})] :\ \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}},\ \mathcal{W}_p(\boldsymbol{\mathcal{C}},\boldsymbol{\mathcal{C}}^0)\le\theta\right\}. \tag{1.4}$$

1.3 Outline of the Thesis and Main Contributions

In Chapter 2, we consider sets of distributions that are within a chosen Wasserstein distance from a nominal distribution, for example an empirical distribution resulting from available

[1] Some authors consider KL ball centered at some nominal distribution instead of nominal copula. Nevertheless, it can be easily shown that the KL divergence between two distributions equals the KL divergence between their associated copulas (cf. Sec 10.4 in [57]).

data. We point out that such a choice of sets has two advantages: (i) The resulting distributions hedged against are more reasonable than those resulting from other popular choices of sets. In particular, they are suitable for *high-dimensional uncertainty* and *uncertainty due to data perturbation*. (ii) The problem of determining the worst-case expectation over the resulting set of distributions has desirable tractability properties. More specifically,

(i) We derive a dual reformulation of the corresponding Wasserstein DRSO problem by *constructing* approximate worst-case distributions (or an exact worst-case distribution if it exists) explicitly via the first-order optimality conditions of the dual problem. The worst-case distributions have a concise structure and a clear interpretation.

(ii) We identify necessary and sufficient conditions for the existence of a worst-case distribution, which are naturally related to the *growth rate* of the objective function.

(iii) Using the structure of the worst-case distribution, we show that data-driven Wasserstein DRSO problems can be approximated to any accuracy by robust optimization problems, and thereby many Wasserstein DRSO problems become tractable by using tools from robust optimization.

(iv) Our strong duality result holds in a very general setting (see Section 2.1 for a detail comparison with existing work), and we show that it can be applied to infinite dimensional process control/estimation problems and worst-case value-at-risk analysis.

By the time we completed the first version of [62], we learned that [63] also considered a similar problem to ours and also obtained a strong duality result. Our focus and our approach to this problem differ from theirs in the following ways. First, we prove the strong duality result for the inner maximization of (Wasserstein-DRSO) using a novel, yet simple, *constructive* approach, in contrast with the non-constructive approaches in their work and also in [24] and [25]. This enables us to establish the structural characterization of the worst-case distributions of the data-driven DRSO (Corollary 2.2(ii)), which improves

the result of [22] and a more recent result in [64] on extremal distributions of Wasserstein balls (Remark 2.5). It also enables us to build a close connection between DRSO and robust optimization (Corollary 2.2(iii)). Second, we focus on Wasserstein distance of order p ($p \geq 1$), while they consider more general transport metrics in which the distance between two points $\xi, \xi' \in \Xi$ is measured by a lower semicontinuous function rather than a metric $\mathsf{d}^p(\xi, \xi')$ as in our case. Nevertheless, our proof remains valid for such more general transport metrics (Remark 2.2). In the meantime, focusing on Wasserstein distance enables us to relate the condition for the existence of a worst-case distribution to the important notion of the "growth rate" of the objective function, and enables us to provide practical guidance for choosing the ambiguity set and controlling the degree of conservativeness based on the objective function (Remark 2.1).

In Chapter 3, we establish a connection between Wasserstein DRSO and regularization. Specifically,

(i) For linear function class with Lipschitz loss, we show an equivalence between parameter-norm regularization and (Wasserstein-DRSO) with $p = 1$. Comparing to [43], we drop the convexity assumption of the loss function, nor do we need any assumption on the data distribution (such as non-separability for SVM as specified in [39]). Instead, we require certain conditions on the asymptotics of the loss function, which is satisfied by many statistical learning problems.

(ii) In the special case of linear optimization, we prove a general equivalence between regularization and (Wasserstein-DRSO) with $p = 1$, allowing arbitrary nominal distribution (not only the empirical one) and a more general metric structure on the data space. Such equivalence has interesting implication in discrete choice modeling – it provides a new interpretation of the discrete choice models from the perspective of distributional robustness, and offers a new economic intuition for the generalized extreme value choice models, which was introduced in the literature pure mathematically.

(iii) For smooth loss function class, we establish an asymptotic equivalence between (Wasserstein-DRSO) with any $p \in [1, \infty]$ and gradient-norm penalty regularization. Such connection suggests a principled way to regularize high-dimensional, non-convex problems, which is demonstrated through the training of Wasserstein generative adversarial networks (WGANs) in deep learning and the estimation of mixed logit model.

In Chapter 4, we investigate distributionally robust optimization with know marginal distributions. We point out that existing studies hedge a family of distributions with known marginals, but either allow arbitrary dependence structure of these distributions, which tends to be over-conservative, or impose constraints on the deviation — measured by Kullback-Leibler divergence — of the dependence structure from some nominal model, which may lead to pathological worst-case distributions. We propose a distributionally robust approach, which hedges against a family of joint distributions with known marginals and a dependence structure similar to — with similarity measured by Wasserstein distance — that of a nominal joint distribution (e.g., the empirical distribution or the independent product distribution). Similarity of the dependence structure is measured through the Wasserstein distance between the copula of the joint distribution and the copula of the nominal distribution. We show that our choice of distance can be used as a new measure of dependence among random variables. Tractability of our new formulation is obtained by a novel constructive proof of strong duality, combining ideas from the theory of multi-marginal optimal transport, variational analysis and a size reduction argument. Numerical experiments in portfolio selection and nonparametric density estimation demonstrate how the proposed approach outperforms other benchmark approaches.

CHAPTER 2
DISTRIBUTIONALLY ROBUST STOCHASTIC OPTIMIZATION WITH WASSERSTEIN DISTANCE

2.1 Overview

This chapter is based on [62].

We first motivate our choice of Wasserstein distance in Section 2.2 by taking close look at the potential issues of divergence measures using examples from computer vision. In Section 2.3, we review some results on the Wasserstein distance.

Next in Section 2.4, we show the tractability of DRSO with Wasserstein distance by developing a strong dual reformulation based on a novel constructive proof. We prove a strong duality result in a very general setting. We show that

$$\sup_{\boldsymbol{\mu}\in\mathcal{P}(\Xi)} \left\{\mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta,\boldsymbol{\xi})] \;:\; \mathcal{W}_p(\boldsymbol{\mu},\boldsymbol{\nu}) \le \theta\right\} \;=\; \min_{\lambda\ge 0}\left\{\lambda\theta^p - \int_{\Xi} \inf_{\xi\in\Xi}[\lambda \mathsf{d}^p(\xi,\zeta) - \ell(\beta,\xi)]\boldsymbol{\nu}(d\zeta)\right\}$$

holds for any Polish space (Ξ, d) and measurable function ℓ (Theorem 2.1). In comparison,

1. Both [24] and [25] assume that Ξ is a convex subset of $\mathbb{R}^K$ with some associated norm. The greater generality of our results enables one to consider interesting problems such as the process control problems in Sections 2.5.1 and 2.5.2, where Ξ is the set of finite counting measures on $[0, 1]$, which is infinite-dimensional and non-convex.

2. Both [24] and [25] assume that the nominal distribution $\boldsymbol{\nu}$ is an empirical distribution, while we allow $\boldsymbol{\nu}$ to be any Borel probability measure. The greater generality enables one to study problems such as the worst-case Value-at-Risk analysis in Section 2.5.3.

3. Both [24] and [25] only consider Wasserstein distance of order $p = 1$. By consider-

ing a bigger family of Wasserstein distances, we establish the importance for DRSO problems of the notion of the "growth rate" of the objective function, which measures how fast the objective function grows compared to a polynomial of order p. It turns out that the growth rate of the objective function determines the finiteness of the worst-case objective value (Proposition 2.2), and it plays an important role in the existence conditions for the worst-case distribution (Corollary 2.1). This is of practical importance, since it provides guidance for choosing the proper Wasserstein distance and for controlling the degree of conservativeness based on the structure of the objective function.

We prove the strong duality result using a novel, elementary, constructive approach. The results of [24] and [25] and other strong duality results in the literature are based on the established Hahn-Banach theorem for certain infinite dimensional vector spaces. In contrast, our proof idea is new and is relatively elementary and straightforward: we use the weak duality result as well as the first-order optimality condition of the dual problem to construct a sequence of primal feasible solutions whose objective values converge to the dual optimal value. Our proof uses relatively elementary tools, without resorting to other "big hammers".

This approach reveals the concise yet insightful structure of the worst-case distribution. As a by product of our constructive proof, we identify necessary and sufficient conditions for the existence of worst-case distributions, and a structural characterization of worst-case distributions (Corollary 2.1). Specifically, for data-driven DRSO problems where $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{\widehat{\xi}^i}$ (where $\boldsymbol{\delta}_\xi$ denotes the unit mass on ξ), whenever a worst-case distribution exists, there is a worst-case distribution $\boldsymbol{\mu}^*$ supported on at most $n+1$ points with the following concise structure:

$$\boldsymbol{\mu}^* \;=\; \frac{1}{n}\sum_{\substack{i=1\\ i\neq i_0}}^{n} \boldsymbol{\delta}_{\xi_*^i} + \frac{p_0}{n}\boldsymbol{\delta}_{\underline{\xi}_*^{i_0}} + \frac{1-p_0}{n}\boldsymbol{\delta}_{\bar{\xi}_*^{i_0}},$$

for some $i_0 \in \{1, \ldots, n\}$, $p_0 \in [0, 1]$ and

$$\xi_*^i \in \operatorname*{arg\,min}_{\xi \in \Xi} \left\{ \lambda^* \mathsf{d}^p(\xi, \widehat{\xi}^i) - \ell(\beta, \xi) \right\}, \ \forall\, i \neq i_0, \ \ \underline{\xi}_*^{i_0}, \overline{\xi}_*^{i_0} \in \operatorname*{arg\,min}_{\xi \in \Xi} \left\{ \lambda^* \mathsf{d}^p(\xi, \widehat{\xi}^{i_0}) - \ell(\beta, \xi) \right\},$$

where λ^* is the dual minimizer (Corollary 2.2). Thus $\boldsymbol{\mu}^*$ can be viewed as a *perturbation* of $\boldsymbol{\nu}$, where the mass on $\widehat{\xi}^i$ is perturbed to ξ_*^i for all $i \neq i_0$, a fraction p_0 of the mass on $\widehat{\xi}^{i_0}$ is perturbed to $\underline{\xi}_*^{i_0}$, and the remaining fraction $1 - p_0$ of the mass on $\widehat{\xi}^{i_0}$ is perturbed to $\overline{\xi}_*^{i_0}$. In particular, uncertainty quantification problems have a worst-case distribution with this simple structure, and can be solved by a greedy procedure (Example 2.7). Our result regarding the existence of a worst-case distribution with such a structure improves the result of [22] and the more recent result of [64] regarding the extremal distributions of Wasserstein balls.

Moreover, a deeper understanding of the worst-case distribution enables us to establish a close connection between Wasserstein DRSO and the traditional robust optimization. Using the structure of a worst-case distribution, we prove that data-driven DRSO problems can be approximated by robust optimization problems to any accuracy (Corollary 2.2(iii)). We use this result to show that two-stage linear DRSO problems with linear decision rules have a tractable semi-definite programming approximation (Section 2.6.2). Moreover, the robust optimization approximation becomes exact when the objective function ℓ is concave in ξ. In addition, if ℓ is convex in β, then the corresponding DRSO problem can be formulated as a convex-concave saddle point problem.

Finally, in Sections 2.5 and 2.6, we apply our results on strong duality and the structural description of the worst-case distributions to a variety of DRSO problems. We conclude this chapter in Section 2.7. Auxiliary results, as well as proofs of some Lemmas, Corollaries and Propositions, are provided in the Appendix A.

2.2 Motivation: potential issues of divergence measures

Despite its widespread use, ϕ-divergence has a number of shortcomings. Here we highlight some of these shortcomings. In a typical setup using ϕ-divergence, Ξ is partitioned into $\bar{B}+1$ bins represented by points $\xi^0, \xi^1, \ldots, \xi^{\bar{B}} \in \Xi$. The nominal distribution $\boldsymbol{\nu}$ associates n_i observations with bin i. That is, the nominal distribution is given by $\boldsymbol{\nu} := (n_0/n, n_1/n, \ldots, n_{\bar{B}}/n)$, where $n := \sum_{i=0}^{\bar{B}} n_i$. Let $\Delta_{\bar{B}} := \{(p_0, p_1, \ldots, p_{\bar{B}}) \in \mathbb{R}_+^{\bar{B}+1} : \sum_{j=0}^{\bar{B}} p_j = 1\}$ denote the set of probability distributions on the same set of bins. Let $\phi : [0, \infty) \mapsto \mathbb{R}$ be a chosen convex function such that $\phi(1) = 0$, with the conventions that $0\phi(a/0) := a \lim_{t\to\infty} \phi(t)/t$ for all $a > 0$, and $0\phi(0/0) := 0$. Then the ϕ-divergence between $\boldsymbol{\mu} = (p_0, \ldots, p_B), \boldsymbol{\nu} = (q_0, \ldots, q_B) \in \Delta_{\bar{B}}$ is defined by

$$I_\phi(\boldsymbol{\mu}, \boldsymbol{\nu}) \;:=\; \sum_{j=0}^{\bar{B}} q_j \phi\left(\frac{p_j}{q_j}\right).$$

Let $\theta > 0$ denote a chosen radius. Then $\mathfrak{M}_\phi := \{\boldsymbol{\mu} \in \Delta_{\bar{B}} : I_\phi(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \theta\}$ denotes the set of probability distributions given by the chosen ϕ-divergence and radius θ. The DRSO problem corresponding to the ϕ-divergence ball $\mathfrak{M}_\phi$ is then given by

$$\inf_{\beta\in\mathcal{D}} \sup_{\boldsymbol{\mu}\in\Delta_{\bar{B}}} \left\{ \sum_{j=0}^{\bar{B}} p_j \ell(\beta, \xi^j) \;:\; I_\phi(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \theta \right\}.$$

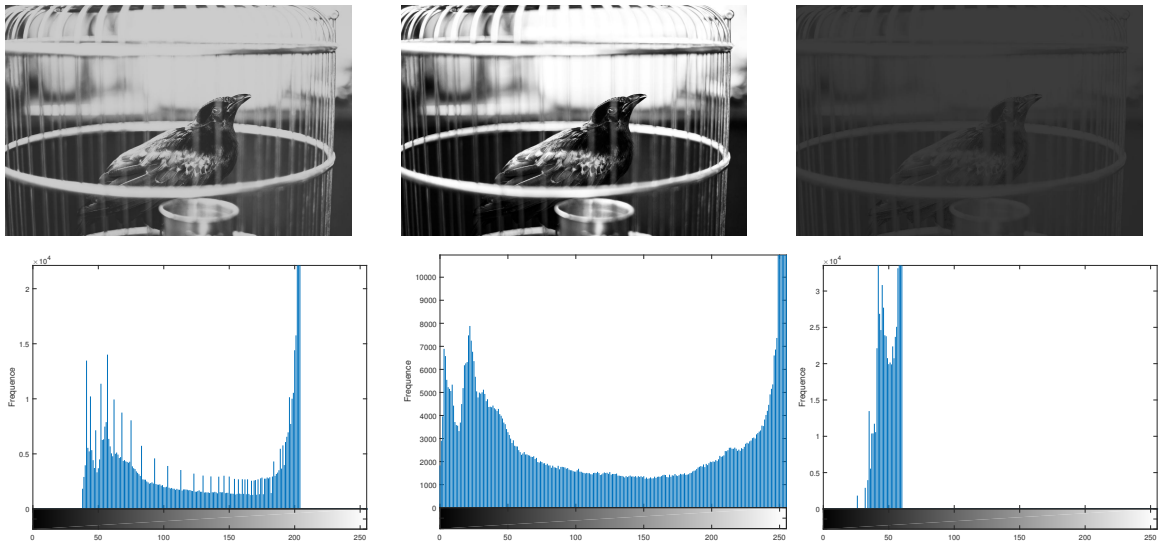
It has been shown in [18] that the ϕ-divergence ball $\mathfrak{M}_\phi$ can be viewed as a statistical confidence region [65], and for several choices of ϕ, the inner maximization of the problem above is tractable.

One well-known shortcoming of ϕ-divergence balls is that, either they are not rich enough to contain distributions that are often relevant, or they they hedge against many distributions that are too extreme. For example, for some choices of ϕ-divergence such as Kullback-Leibler divergence, if the nominal $q_i = 0$, then $p_i = 0$, that is, the ϕ-divergence ball includes only distributions that are absolutely continuous with respect to the nominal

distribution $\boldsymbol{\nu}$, and thus does not include distributions with support on points where the nominal distribution $\boldsymbol{\nu}$ is not supported. As a result, if $\Xi = \mathbb{R}^K$ and $\boldsymbol{\nu}$ is discrete, then there are no continuous distributions in the ϕ-divergence ball $\mathfrak{M}_\phi$. Some other choices of ϕ-divergence exhibit in some sense the opposite behavior. For example, the Burg entropy ball includes distributions with some amount of probability allowed to be shifted from $\boldsymbol{\nu}$ to any other bin, with the amount of probability allowed to be shifted depending only on θ and not on how extreme the bin is. See Section 2.6.1 for more details regarding this potential shortcoming. Moreover, for two high-dimensional distributions supported on two low-dimensional manfiolds with measure-zero intersection, their ϕ-divergence will be maxed out (e.g., in KL case it equals $+\infty$), and thus ϕ-divergence is not a good measure of closeness between high-dimensional distributions with low-dimensional structure [66].

Next we illustrate another shortcoming of ϕ-divergence that will motivate the use of Wasserstein distance.

Example 2.1. Suppose that there is an underlying true image (2.1b), and a decision maker possesses, instead of the true image, an approximate image (2.1a) obtained with a less than perfect device that loses some of the contrast. The images are summarized by their gray-scale histograms. (In fact, (2.1a) was obtained from (2.1b) by a low-contrast intensity transformation [67], by which the black pixels become somewhat whiter and the white pixels become somewhat blacker. This type of transformation changes only the gray-scale value of a pixel and not the location of a gray-scale value, and therefore it can also be regarded as a transformation from one gray-scale histogram to another gray-scale histogram.) As a result, roughly speaking, the observed histogram $\boldsymbol{\nu}$ is obtained by shifting the true histogram $\boldsymbol{\mu}_{true}$ inwards. Also consider the pathological image (2.1c) that is too dark to see many details, with histogram $\boldsymbol{\mu}_{pathol}$. Suppose that the decision maker constructs a Kullback-Leibler (KL) divergence ball $\mathfrak{M}_{\phi_{KL}} := \{\boldsymbol{\mu} \in \Delta_{\bar{B}} \,:\, I_{\phi_{KL}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \theta\}$. Note that $I_{\phi_{KL}}(\boldsymbol{\mu}_{true}, \boldsymbol{\nu}) = 5.05 > I_{\phi_{KL}}(\boldsymbol{\mu}_{pathol}, \boldsymbol{\nu}) = 2.33$. Therefore, if θ is chosen small enough (less than 2.33) for $\mathfrak{M}_{\phi_{KL}}$ to exclude the pathological image (2.1c), then $\mathfrak{M}_{\phi_{KL}}$ will also



(a) Observed image with histogram $\boldsymbol{\nu}$

(b) True image with histogram $\boldsymbol{\mu}_{true}$

(c) Pathological image with histogram $\boldsymbol{\mu}_{pathol}$

Figure 2.1: Three images and their gray-scale histograms.

exclude the true image (2.1b). If θ is chosen large enough (greater than 5.05) for $\mathfrak{M}_{\phi_{KL}}$ to include the true image (2.1b), then $\mathfrak{M}_{\phi_{KL}}$ also has to include the pathological image (2.1c), and then the resulting decision may be overly conservative due to hedging against irrelevant distributions. If an intermediate value is chosen for θ (between 2.33 and 5.05), then $\mathfrak{M}_{\phi_{KL}}$ includes the pathological image (2.1c) and excludes the true image (2.1b). In contrast, note that the Wasserstein distance $\mathcal{W}_1$ satisfies $\mathcal{W}_1(\boldsymbol{\mu}_{true}, \boldsymbol{\nu}) = 30.7 < \mathcal{W}_1(\boldsymbol{\mu}_{pathol}, \boldsymbol{\nu}) = 84.0$, and thus Wasserstein distance does not exhibit the problem encountered with KL divergence (see also Example 2.3).

The reason for such behavior is that ϕ-divergence does not incorporate a notion of how close two points $\xi, \xi' \in \Xi$ are to each other, for example, how likely it is that observation is ξ' given that the true value is ξ. In Example 2.1, $\Xi = \{0, 1, \ldots, 255\}$ represents 8-bit gray-scale values. In this case, we know that the likelihood that a pixel with gray-scale value $\xi \in \Xi$ is observed with gray-scale value $\xi' \in \Xi$ is decreasing in the absolute difference between ξ and ξ'. However, in the definition of ϕ-divergence, only the *relative ratio* p_j/q_j for the same gray-scale value j is taken into account, while the distances between different

gray-scale values is not taken into account. This phenomenon has been observed in studies of image retrieval [68, 69].

The drawbacks of ϕ-divergence motivates us to consider sets $\mathfrak{M}$ that incorporate a notion of how close two points $\xi, \xi' \in \Xi$ are to each other. One such choice of $\mathfrak{M}$ is based on Wasserstein distance.

2.3 Notation and Preliminaries

In this section, we introduce notation and briefly outline some known results regarding Wasserstein distance. For a more detailed discussion we refer to [70, 71].

Let Ξ be a Polish (separable complete metric) space with metric d. Let $\mathscr{B}(\Xi)$ denote the Borel σ-algebra on Ξ, and let $\mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ denote the *completion* of $\mathscr{B}(\Xi)$ with respect to a measure $\boldsymbol{\nu}$ on $\mathscr{B}(\Xi)$ such that the measure space $(\Xi, \mathscr{B}_{\boldsymbol{\nu}}(\Xi), \boldsymbol{\nu})$ is complete (see, e.g., Definition 1.11 in [72]). Let $\mathcal{B}(\Xi)$ denote the set of Borel measures on Ξ, let $\mathcal{P}(\Xi)$ denote the set of Borel probability measures on Ξ, and let $\mathcal{P}_p(\Xi)$ denote the subset of $\mathcal{P}(\Xi)$ with finite p-th moment for $p \in [1, \infty)$:

$$\mathcal{P}_p(\Xi) \;:=\; \left\{ \boldsymbol{\mu} \in \mathcal{P}(\Xi) \;:\; \int_{\Xi} \mathsf{d}^p(\xi, \zeta^0) \boldsymbol{\mu}(d\xi) \;<\; \infty \text{ for some } \zeta^0 \in \Xi \right\}.$$

It follows from the triangle inequality that the definition above does not depend on the choice of ζ^0. A function $\ell : \Xi \to \mathbb{R}$ is called $\boldsymbol{\nu}$-measurable if it is $(\mathscr{B}_{\boldsymbol{\nu}}(\Xi), \mathscr{B}(\mathbb{R}))$-measurable, and a function $T : \Xi \mapsto \Xi$ is called $\boldsymbol{\nu}$-measurable if it is $(\mathscr{B}_{\boldsymbol{\nu}}(\Xi), \mathscr{B}(\Xi))$-measurable. To facilitate later discussion, we introduce the push-forward operator on measures.

Definition 2.1 (Push-forward Measure)**.** Given measurable spaces $(\Xi, \mathscr{B}(\Xi))$ and $(\Xi', \mathscr{B}(\Xi'))$, a measurable function $T : \Xi \mapsto \Xi'$, and a measure $\boldsymbol{\nu} \in \mathcal{B}(\Xi)$, let $T_{\#}\boldsymbol{\nu} \in \mathcal{B}(\Xi')$ denote the

push-forward measure of $\boldsymbol{\nu}$ through T, defined by

$$T_{\#}\boldsymbol{\nu}(A) \;:=\; \boldsymbol{\nu}(T^{-1}(A)) \;=\; \boldsymbol{\nu}\{\zeta \in \Xi \,:\, T(\zeta) \in A\},\ \forall \text{ measurable sets } A \subset \Xi'.$$

That is, $T_{\#}\boldsymbol{\nu}$ is obtained by *transporting* ("pushing forward") $\boldsymbol{\nu}$ from Ξ to Ξ' using the function T. For $i \in \{1, 2\}$, let $\pi^i : \Xi \times \Xi \mapsto \Xi$ denote the canonical projections given by $\pi^i(\xi^1, \xi^2) = \xi^i$. Then for a measure $\boldsymbol{\gamma} \in \mathcal{P}(\Xi \times \Xi)$, $\pi^i_{\#}\boldsymbol{\gamma} \in \mathcal{P}(\Xi)$ is the i-th marginal of $\boldsymbol{\gamma}$ given by $\pi^1_{\#}\boldsymbol{\gamma}(A) = \boldsymbol{\gamma}(A \times \Xi)$ and $\pi^2_{\#}\boldsymbol{\gamma}(A) = \boldsymbol{\gamma}(\Xi \times A)$.

Definition 2.2 (p-Wasserstein distance)**.** The Wasserstein distance $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$ between $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_p(\Xi)$ is defined by

$$\mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \;:=\; \min_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi \times \Xi)} \left\{ \int_{\Xi \times \Xi} \mathsf{d}^p(\xi, \zeta) \boldsymbol{\gamma}(d\xi, d\zeta) \;:\; \pi^1_{\#}\boldsymbol{\gamma} = \boldsymbol{\mu}, \pi^2_{\#}\boldsymbol{\gamma} = \boldsymbol{\nu} \right\}. \qquad (2.1)$$

That is, the Wasserstein distance between $\boldsymbol{\mu}, \boldsymbol{\nu}$ is the minimum cost (in terms of d^p) of redistributing mass from $\boldsymbol{\nu}$ to $\boldsymbol{\mu}$, which is why it is also called the "earth mover's distance". Wasserstein distance is a natural way of comparing two distributions when one is obtained from the other by *perturbations*. The minimum on the right side of (4.1) is attained, because d is non-negative, continuous and thus lower semicontinuous (Theorem 1.3 of [70]). The following example is a familiar special case of problem (4.1).

Example 2.2 (Transportation problem)**.** Consider $\boldsymbol{\mu} = \sum_{i=1}^m p_i \boldsymbol{\delta}_{\xi^i}$ and $\boldsymbol{\nu} = \sum_{j=1}^n q_j \boldsymbol{\delta}_{\widehat{\xi}^j}$, where $m, n \geq 1$, $p_i, q_j \geq 0$, $\xi^i, \widehat{\xi}^j \in \Xi$ for all i, j, and $\sum_{i=1}^m p_i = \sum_{j=1}^n q_j = 1$. Then problem (4.1) becomes the classical transportation problem in linear programming:

$$\min_{\boldsymbol{\gamma}_{ij} \geq 0} \left\{ \sum_{i=1}^m \sum_{j=1}^n \mathsf{d}^p(\xi^i, \widehat{\xi}^j) \boldsymbol{\gamma}_{ij} \;:\; \sum_{j=1}^n \boldsymbol{\gamma}_{ij} = p_i,\ \forall\, i,\ \sum_{i=1}^m \boldsymbol{\gamma}_{ij} = q_j,\ \forall\, j \right\}.$$

Example 2.3 (Revisiting Example 2.1)**.** Next we evaluate the Wasserstein distance between the histograms in Example 2.1. To evaluate $\mathcal{W}_1(\boldsymbol{\mu}_{true}, \boldsymbol{\nu})$, note that the least cost way of transporting mass from $\boldsymbol{\nu}$ to $\boldsymbol{\mu}_{true}$ is to move the mass outwards. In contrast, to evalu-

ate $\mathcal{W}_1(\boldsymbol{\mu}_{pathol}, \boldsymbol{\nu})$, one has to transport mass relatively long distances from right to left (changing the gray-scale values of pixels by large amounts), resulting in a larger cost than $\mathcal{W}_1(\boldsymbol{\mu}_{true}, \boldsymbol{\nu})$. Therefore $\mathcal{W}_1(\boldsymbol{\mu}_{pathol}, \boldsymbol{\nu}) > \mathcal{W}_1(\boldsymbol{\mu}_{true}, \boldsymbol{\nu})$.

Wasserstein distance has a dual representation due to Kantorovich's duality (Theorem 5.10 in [71]):

$$\begin{aligned} &\mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \\ &= \sup_{u \in L^1(\boldsymbol{\mu}), v \in L^1(\boldsymbol{\nu})} \left\{ \int_\Xi u(\xi) \boldsymbol{\mu}(d\xi) + \int_\Xi v(\zeta) \boldsymbol{\nu}(d\zeta) : \ u(\xi) + v(\zeta) \leq \mathsf{d}^p(\xi, \zeta), \forall \xi, \zeta \in \Xi \right\}, \end{aligned} \tag{2.2}$$

where $L^1(\boldsymbol{\nu})$ represents the L^1 space of $\boldsymbol{\nu}$-measurable functions. In addition, the set of functions under the supremum above can be replaced by $u, v \in C_b(\Xi)$, where $C_b(\Xi)$ denotes the set of continuous and bounded real-valued functions on Ξ. Particularly, when $p = 1$, by the Kantorovich-Rubinstein Theorem, (2.2) can be simplified to (see, e.g., Equation (5.11) in [71])

$$\mathcal{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}) \ = \ \sup_{u \in L^1(\boldsymbol{\mu})} \left\{ \int_\Xi u(\xi) d(\boldsymbol{\mu} - \boldsymbol{\nu})(\xi) \ : \ u \text{ is 1-Lipschitz} \right\}.$$

So for an L-Lipschitz function $\ell : \Xi \mapsto \mathbb{R}$, it holds that $\left| \mathbb{E}_{\boldsymbol{\mu}}[\ell(\xi)] - \mathbb{E}_{\boldsymbol{\nu}}[\ell(\xi)] \right| \leq L\mathcal{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq L\theta$ for all $\boldsymbol{\mu} \in \mathfrak{M}$.

We remark that Definition 4.2 and the results above can be extended to finite Borel measures. Moreover, we have the following result.

Lemma 2.1. *For any finite Borel measures $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{B}(\Xi)$ with $\boldsymbol{\mu}(\Xi) \neq \boldsymbol{\nu}(\Xi)$, it holds that $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \infty$.*

Another important feature of Wasserstein distance is that $\mathcal{W}_p$ metrizes weak convergence in $\mathcal{P}_p(\Xi)$ (cf. Theorem 6.9 in [71]). That is, for any sequence $\{\boldsymbol{\mu}_k\}_{k=1}^\infty$ of measures in $\mathcal{P}_p(\Xi)$ and $\boldsymbol{\mu} \in \mathcal{P}_p(\Xi)$, it holds that $\lim_{k \to \infty} \mathcal{W}_p(\boldsymbol{\mu}_k, \boldsymbol{\mu}) = 0$ if and only if $\boldsymbol{\mu}_k$ converges weakly to $\boldsymbol{\mu}$ and $\int_\Xi \mathsf{d}^p(\xi, \zeta^0) \boldsymbol{\mu}_k(d\xi) \to \int_\Xi \mathsf{d}^p(\xi, \zeta^0) \boldsymbol{\mu}(d\xi)$ as $k \to \infty$. Therefore, conver-

gence in the Wasserstein distance of order p implies convergence up to the p-th moment. [71, chapter 6] discusses the advantages of Wasserstein distance relative to other distances, such as the Prokhorov metric, that metrize weak convergence.

2.4 Tractable Reformulation via Duality

In this section we develop a tractable reformulation by deriving its strong dual. We suppress the variable β of ℓ in this section, and results are interpreted pointwise for each β. Given $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$ and $\ell \in L^1(\boldsymbol{\nu})$, for any $\theta > 0$ and $p \in [1, \infty)$, the inner maximization problem of (Wasserstein-DRSO) is written as

$$v_P \;:=\; \sup_{\boldsymbol{\mu} \in \mathfrak{M}} \int_\Xi \ell(\xi) \boldsymbol{\mu}(d\xi) \;=\; \sup_{\boldsymbol{\mu} \in \mathcal{P}(\Xi)} \left\{ \int_\Xi \ell(\xi) \boldsymbol{\mu}(d\xi) \;:\; \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \theta \right\}. \quad \text{(Primal)}$$

Our main goal is to derive its strong dual

$$v_D \;:=\; \inf_{\lambda \ge 0} \left\{ \lambda \theta^p - \int_\Xi \inf_{\xi \in \Xi} \left[\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \right] \boldsymbol{\nu}(d\zeta) \right\}. \quad \text{(Dual)}$$

The dual problem is a one-dimensional convex minimization problem with respect to λ, the Lagrangian multiplier of the Wasserstein constraint in the primal problem. The term $\inf_{\xi \in \Xi}[\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)]$ is called the *Moreau-Yosida regularization* of $-\ell$ with parameter $1/\lambda$ in the literature (cf. [73]). Its measurability with respect to $\boldsymbol{\nu}$ is established in Lemma 2.4(i) in Section 2.4.1.

2.4.1 General Nominal Distribution

In this section, we prove the strong duality result for a general nominal distribution $\boldsymbol{\nu}$ on a Polish space Ξ. Such generality broadens the applicability of the result for (Wasserstein-DRSO). For example, the result is useful when the nominal distribution is some distribution such as a Gaussian distribution on $\mathbb{R}^K$ (Section 2.5.3), or even some stochastic process (Sections 2.5.1 and 2.5.2). We begin with a weak duality result, which is an application of

Lagrangian weak duality.

Proposition 2.1 (Weak duality)**.** *Consider any $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$ and $\ell \in L^1(\boldsymbol{\nu})$. Then for any $p \in [1, \infty)$ and $\theta > 0$, it holds that $v_P \leq v_D$.*

To prove the strong duality result, we consider two separate cases: $v_D = \infty$ and $v_D < \infty$. As can be seen from (Dual), if the term $-\int_\Xi \inf_{\xi \in \Xi} \left[\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\right] \boldsymbol{\nu}(d\zeta)$ is infinite for all $\lambda \geq 0$, then $v_D = \infty$. Thus, to facilitate our analysis, we introduce the following definitions.

Definition 2.3 (Regularization Operator Φ)**.** Let $\Phi : \mathbb{R} \times \Xi \to \mathbb{R} \cup \{-\infty\}$ be given by

$$\Phi(\lambda, \zeta) \;:=\; \inf_{\xi \in \Xi} \left\{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\right\}.$$

For any $\lambda \geq 0$ and any $\zeta \in \Xi$ such that $\Phi(\lambda, \zeta) > -\infty$, let

$$\begin{aligned} \overline{D}(\lambda, \zeta) \;&:=\; \limsup_{\varepsilon \downarrow 0} \left\{ \sup_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) \;:\; \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \;\leq\; \Phi(\lambda, \zeta) + \varepsilon \right\} \right\}, \\ \underline{D}(\lambda, \zeta) \;&:=\; \liminf_{\varepsilon \downarrow 0} \left\{ \inf_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) \;:\; \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \;\leq\; \Phi(\lambda, \zeta) + \varepsilon \right\} \right\}. \end{aligned} \tag{2.3}$$

For any $\lambda \geq 0$ and any $\zeta \in \Xi$ such that $\arg\min_{\xi \in \Xi}\{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\}$ is nonempty, let

$$\begin{aligned} \overline{D}_0(\lambda, \zeta) \;&:=\; \sup_{\xi \in \Xi} \{ \mathsf{d}^p(\xi, \zeta) \;:\; \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) = \Phi(\lambda, \zeta) \} \\ \underline{D}_0(\lambda, \zeta) \;&:=\; \inf_{\xi \in \Xi} \{ \mathsf{d}^p(\xi, \zeta) \;:\; \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) = \Phi(\lambda, \zeta) \} \end{aligned} \tag{2.4}$$

Then $\underline{D}_0(\lambda, \zeta)$ and $\overline{D}_0(\lambda, \zeta)$ represent respectively the closest and furthest distances between ζ and any point in $\arg\min_{\xi \in \Xi}\{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\}$. Note that $\overline{D}_0(\lambda, \zeta)$ (resp. $\underline{D}_0(\lambda, \zeta)$) may not be equal to $\overline{D}(\lambda, \zeta)$ (resp. $\underline{D}(\lambda, \zeta)$).

Definition 2.4 (Growth rate)**.** Define the *growth rate* κ of ℓ as

$$\kappa \;:=\; \inf \left\{ \lambda \geq 0 \;:\; \int_\Xi \Phi(\lambda, \zeta) \boldsymbol{\nu}(d\zeta) > -\infty \right\}.$$

Particularly, if $\int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) = -\infty$ for all $\lambda \geq 0$, then $\kappa = \infty$.

If Ξ is bounded and ℓ is bounded above, then $\kappa = 0$, and if Ξ is bounded and ℓ is not bounded above, then $\kappa = \infty$. The possibilities when Ξ is not bounded are more interesting. Next, Lemma 2.2 establishes some additional properties of κ, including the property that if Ξ is not bounded and $\kappa < \infty$, then

$$\kappa \;=\; \limsup_{\xi \in \Xi \,:\, \mathsf{d}^p(\xi,\zeta) \to \infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta)\}}{\mathsf{d}^p(\xi, \zeta)}$$

for any $\zeta \in \Xi$, which motivates why we call κ the growth rate of ℓ.

Lemma 2.2 (Properties of the growth rate κ)**.**

(i) Suppose that Ξ is unbounded. Then the quantity

$$\limsup_{\xi \in \Xi \,:\, \mathsf{d}^p(\xi,\zeta) \to \infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta)\}}{\mathsf{d}^p(\xi, \zeta)}$$

is independent of ζ.

(ii) Suppose that $\boldsymbol{\nu} \in \mathcal{P}_p(\Xi)$. Then the growth rate κ is finite if and only if there exists $\zeta^0 \in \Xi$ and $L, M > 0$ such that

$$\ell(\xi) - \ell(\zeta^0) \;\leq\; L\mathsf{d}^p(\xi, \zeta^0) + M \quad \forall\, \xi \in \Xi. \tag{2.5}$$

(iii) Suppose that $\boldsymbol{\nu} \in \mathcal{P}_p(\Xi)$. If Ξ is unbounded and $\kappa < \infty$, then

$$\kappa \;=\; \limsup_{\xi \in \Xi \,:\, \mathsf{d}^p(\xi,\zeta) \to \infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta)\}}{\mathsf{d}^p(\xi, \zeta)}$$

for any $\zeta \in \Xi$.

Remark 2.1 (Choosing Wasserstein order p)**.** Let

$$\underline{p} \;:=\; \inf\Big\{p \geq 1 \;:\; \limsup_{d(\zeta,\zeta^0)\to\infty} \frac{\ell(\zeta) - \ell(\zeta^0)}{\mathsf{d}^p(\zeta,\zeta^0)} < \infty\Big\}.$$

Proposition 2.2 suggests that a meaningful formulation of (Wasserstein-DRSO) should be such that the Wasserstein order p is greater than or equal to $\underline{p}$. In both [24] and [25] only $p = 1$ is considered. By considering higher orders p in our analysis, we can accommodate a greater set of functions ℓ, and we also have more flexibility to choose the ambiguity set and to control the degree of conservativeness.

Lemma 2.3 (Properties of the regularization operator Φ)**.** *Let (Ξ, d) be a Polish space. Consider any $p \in [1,\infty)$, $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, and $\ell \in L^1(\boldsymbol{\nu})$ such that $\kappa < \infty$. Then there is a set $B \in \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ such that $\boldsymbol{\nu}(B) = 1$, and the following holds.*

(i) [Monotonicity] $\Phi(\cdot,\zeta)$ is nondecreasing and upper-semicontinuous for all $\zeta \in \Xi$. $\Phi(\lambda,\zeta) > -\infty$ for all $\lambda > \kappa$ and all $\zeta \in B$. $\Phi(\cdot,\zeta)$ is concave for all $\zeta \in B$. For any $\lambda_2 > \lambda_1$ such that $\Phi(\lambda_1,\zeta) > -\infty$, it holds that $\overline{D}(\lambda_2,\zeta) \leq \underline{D}(\lambda_1,\zeta) \leq \overline{D}(\lambda_1,\zeta)$.

(ii) [Bounds] For any $\lambda_2 > \lambda_1$ such that $\Phi(\lambda_1,\zeta) > -\infty$, it holds that

$$(\lambda_2 - \lambda_1)\overline{D}(\lambda_2,\zeta) \;\;\leq\;\; -\ell(\zeta) - \Phi(\lambda_1,\zeta).$$

(iii) [Derivative] For all $\lambda > \kappa$ and all $\zeta \in B$, the left partial derivative $\partial\Phi(\lambda,\zeta)/\partial\lambda-$ exists and satisfies

$$\overline{D}(\lambda,\zeta) \;\;\leq\;\; \frac{\partial\Phi(\lambda,\zeta)}{\partial\lambda-} \;\;\leq\;\; \lim_{\lambda_1\uparrow\lambda} \underline{D}(\lambda_1,\zeta).$$

For all $\lambda \geq 0$ and $\zeta \in \Xi$ such that $\Phi(\lambda,\zeta) > -\infty$, the right partial derivative

$\partial\Phi(\lambda,\zeta)/\partial\lambda+$ *exists and satisfies*

$$\lim_{\lambda_2\downarrow\lambda}\overline{D}(\lambda_2,\zeta) \;\leq\; \frac{\partial\Phi(\lambda,\zeta)}{\partial\lambda+} \;\leq\; \underline{D}(\lambda,\zeta).$$

Lemma 2.4 (Measurability)**.** *For any $p\in[1,\infty)$, $\boldsymbol{\nu}\in\mathcal{P}(\Xi)$, and $\ell\in L^1(\boldsymbol{\nu})$, the following hold:*

(i) $\Phi(\lambda,\cdot)$, $\overline{D}(\lambda,\cdot)$, $\underline{D}(\lambda,\cdot)$, $\overline{D}_0(\lambda,\cdot)$, *and* $\underline{D}_0(\lambda,\cdot)$ *are* $\boldsymbol{\nu}$*-measurable.*

(ii) Suppose that $\kappa<\infty$. Then for any $\lambda,\delta,\varepsilon\geq 0$ such that the sets

$$\begin{aligned}
\overline{F}(\lambda,\zeta) &:= \Big\{\xi\in\Xi \,:\, \lambda\mathsf{d}^p(\xi,\zeta)-\ell(\xi)\leq\Phi(\lambda,\zeta)+\varepsilon,\ \mathsf{d}^p(\xi,\zeta)\geq\overline{D}(\lambda,\zeta)-\delta\Big\},\\
\underline{F}(\lambda,\zeta) &:= \Big\{\xi\in\Xi \,:\, \lambda\mathsf{d}^p(\xi,\zeta)-\ell(\xi)\leq\Phi(\lambda,\zeta)+\varepsilon,\ \mathsf{d}^p(\xi,\zeta)\leq\underline{D}(\lambda,\zeta)+\delta\Big\}
\end{aligned}$$

are non-empty for $\boldsymbol{\nu}$-almost all $\zeta\in\Xi$, there exists $\boldsymbol{\nu}$-measurable mappings $\overline{T}(\lambda,\cdot),\underline{T}(\lambda,\cdot):\Xi\to\Xi$ such that $\overline{T}(\lambda,\zeta)\in\overline{F}(\lambda,\zeta)$ and $\underline{T}(\lambda,\zeta)\in\underline{F}(\lambda,\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta\in\Xi$.

(iii) Suppose that $\kappa<\infty$. Then for any $\lambda,\delta\geq 0$ such that the sets

$$\begin{aligned}
\overline{F}(\lambda,\zeta) &:= \Big\{\xi\in\Xi \,:\, \lambda\mathsf{d}^p(\xi,\zeta)-\ell(\xi)=\Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta)\geq\overline{D}_0(\lambda,\zeta)-\delta\Big\},\\
\underline{F}(\lambda,\zeta) &:= \Big\{\xi\in\Xi \,:\, \lambda\mathsf{d}^p(\xi,\zeta)-\ell(\xi)=\Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta)\leq\underline{D}_0(\lambda,\zeta)+\delta\Big\}
\end{aligned}$$

are non-empty for $\boldsymbol{\nu}$-almost all $\zeta\in\Xi$, there exists $\boldsymbol{\nu}$-measurable mappings $\overline{T}(\lambda,\cdot),\underline{T}(\lambda,\cdot):\Xi\to\Xi$ such that $\overline{T}(\lambda,\zeta)\in\overline{F}(\lambda,\zeta)$ and $\underline{T}(\lambda,\zeta)\in\underline{F}(\lambda,\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta\in\Xi$.

(iv) For any $E\in\mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ and $a,b>0$ such that

$$F(\zeta) := \{\xi\in\Xi \,:\, \ell(\xi)-\ell(\zeta)>a\mathsf{d}^p(\xi,\zeta)+b\}$$

is non-empty for $\boldsymbol{\nu}$-almost all $\zeta\in E$, there exists a $\boldsymbol{\nu}$-measurable mapping $T:E\to\Xi$ such that $T(\zeta)\in F(\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta\in E$.

(v) Suppose that $\kappa < \infty$. For any $\kappa' \in (0, \kappa)$, and any $\boldsymbol{\nu}$-measurable function $M : \Xi \to \mathbb{R}$ such that the set

$$F(\zeta) \;:=\; \{\xi \in \Xi \,:\, \ell(\xi) - \ell(\zeta) \geq \kappa' \mathsf{d}^p(\xi, \zeta),\ \mathsf{d}^p(\xi, \zeta) \geq M(\zeta)\}$$

is non-empty for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$, there exists a $\boldsymbol{\nu}$-measurable mapping $T : \Xi \to \Xi$ such that $T(\zeta) \in F(\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$.

Let $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ denote the dual objective function given by

$$h(\lambda) \;:=\; \lambda\theta^p - \int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta).$$

Lemma 2.5 (Dual objective function)**.** *The dual objective function h has the following properties:*

(i) $h(\lambda) = \infty$ for all $\lambda \in [0, \kappa)$ and $h(\lambda) < \infty$ for all $\lambda \in (\kappa, \infty)$.

(ii) h is a convex function.

(iii) h is a lower-semicontinuous function.

(iv) $h(\lambda) \to \infty$ as $\lambda \to \infty$.

(v) h has a minimizer $\lambda^* \in [\kappa, \infty)$.

Proposition 2.2 (Strong duality with infinite optimal value)**.** *Consider any $p \in [1, \infty)$, $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, and $\ell \in L^1(\boldsymbol{\nu})$. Suppose that $\theta > 0$ and $\kappa = \infty$. Then $v_P = v_D = \infty$.*

The next theorem establishes a strong duality result when the growth rate κ is finite.

Theorem 2.1 (Strong duality with finite optimal value)**.** *Consider any $p \in [1, \infty)$, any $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, any $\theta > 0$, and any $\ell \in L^1(\boldsymbol{\nu})$ such that $\kappa < \infty$. Then $v_P = v_D < \infty$.*

Proof of Theorem 2.1. Proposition 2.1 established the weak duality result that $v_P \leq v_D$, so it suffices to show that $v_P \geq v_D$. Lemma 2.5(v) established that the dual objective function h has a minimizer $\lambda^* \in [\kappa, \infty)$. Next we consider the two cases (1) h has a minimizer $\lambda^* > \kappa$, or (2) κ is the unique minimizer of h, separately. In each case, we construct a sequence of primal feasible solutions which converges to the dual optimal value by exploiting the first-order optimality condition of the dual.

- Case 1: h has a minimizer $\lambda^* > \kappa$.

Note that for any $\lambda > \kappa$ and $\delta, \varepsilon > 0$, the sets $\overline{F}(\lambda, \zeta), \underline{F}(\lambda, \zeta)$ in Lemma 2.4(ii) are non-empty for all $\zeta \in B$. Hence there exists $\boldsymbol{\nu}$-measurable mappings $\overline{T}(\lambda, \cdot), \underline{T}(\lambda, \cdot) : \Xi \to \Xi$ such that

$$\begin{aligned}
\overline{T}(\lambda, \zeta) &\in \Big\{\xi \in \Xi \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda, \zeta) + \varepsilon,\ \mathsf{d}^p(\xi, \zeta) \geq \overline{D}(\lambda, \zeta) - \delta\Big\}, \\
\underline{T}(\lambda, \zeta) &\in \Big\{\xi \in \Xi \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda, \zeta) + \varepsilon,\ \mathsf{d}^p(\xi, \zeta) \leq \underline{D}(\lambda, \zeta) + \delta\Big\}
\end{aligned}$$

for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$.

The first-order optimality conditions $\frac{\partial}{\partial \lambda -} h(\lambda^*) \leq 0$ and $\frac{\partial}{\partial \lambda +} h(\lambda^*) \geq 0$ imply that

$$\frac{\partial}{\partial \lambda +} \left(\int_\Xi \Phi(\lambda^*, \zeta) \boldsymbol{\nu}(d\zeta) \right) \leq \theta^p \leq \frac{\partial}{\partial \lambda -} \left(\int_\Xi \Phi(\lambda^*, \zeta) \boldsymbol{\nu}(d\zeta) \right). \tag{2.6}$$

Next we verify that we can interchange the partial derivative and integration in (2.6). Recall from Lemma 2.3(i) that there is a set $B \in \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ such that $\boldsymbol{\nu}(B) = 1$, and $\Phi(\lambda, \zeta) > -\infty$ for all $\lambda > \kappa$ and all $\zeta \in B$, and $\Phi(\cdot, \zeta)$ is concave for all $\zeta \in B$. To show it for the right derivative, consider any $\zeta \in B$ and any decreasing sequence $\lambda_n \downarrow \lambda^*$. Let

$$f_n(\zeta) := \frac{\Phi(\lambda_n, \zeta) - \Phi(\lambda^*, \zeta)}{\lambda_n - \lambda^*}$$

Since $\Phi(\cdot, \zeta)$ is nondecreasing for all ζ, $f_n(\zeta) \geq 0$ for all ζ. In addition, since $\Phi(\cdot, \zeta)$ is concave, $f_n \leq f_{n+1}$ for all n. Note that $\lim_{n \to \infty} f_n(\zeta) = \frac{\partial}{\partial \lambda +} \Phi(\lambda^*, \zeta)$. Thus it follows

from the monotone convergence theorem that

$$\frac{\partial}{\partial\lambda+}\left(\int_{\Xi}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_{\Xi}\frac{\partial}{\partial\lambda+}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta). \tag{2.7}$$

To show it for the left derivative in (2.6), consider any $\zeta \in B$ and any increasing sequence $\lambda_n \uparrow \lambda^*$ with $\lambda_1 > \kappa$. Let f_n be defined as before, and thus $f_n(\zeta) \geq 0$ for all ζ. In addition, since $\Phi(\cdot,\zeta)$ is concave, $f_n \geq f_{n+1}$ for all n. That is, for all n it holds that

$$|f_n(\zeta)| \;=\; f_n(\zeta) \;\leq\; f_1(\zeta) \;\leq\; \frac{|\Phi(\lambda_1,\zeta)| + |\Phi(\lambda^*,\zeta)|}{\lambda^* - \lambda_1}$$

It follows from $\lambda_1 > \kappa$ that

$$\int_{\Xi} f_1(\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \frac{\int_{\Xi}|\Phi(\lambda_1,\zeta)|\,\boldsymbol{\nu}(d\zeta) + \int_{\Xi}|\Phi(\lambda^*,\zeta)|\,\boldsymbol{\nu}(d\zeta)}{\lambda^* - \lambda_1} \;<\; \infty$$

Also, $\lim_{n\to\infty} f_n(\zeta) = \frac{\partial}{\partial\lambda-}\Phi(\lambda^*,\zeta)$. Thus it follows from the dominated convergence theorem that

$$\frac{\partial}{\partial\lambda-}\left(\int_{\Xi}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_{\Xi}\frac{\partial}{\partial\lambda-}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta). \tag{2.8}$$

Therefore it follows from (2.6), (2.7), (2.8), and Lemma 2.3(iii) that

$$\begin{aligned}
\theta^p \;&\geq\; \frac{\partial}{\partial\lambda+}\left(\int_{\Xi}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_{\Xi}\frac{\partial}{\partial\lambda+}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) \;\geq\; \int_{\Xi}\lim_{\lambda\downarrow\lambda^*}\overline{D}(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) \\
\theta^p \;&\leq\; \frac{\partial}{\partial\lambda-}\left(\int_{\Xi}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_{\Xi}\frac{\partial}{\partial\lambda-}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \int_{\Xi}\lim_{\lambda\uparrow\lambda^*}\underline{D}(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)
\end{aligned} \tag{2.9}$$

In particular, for any λ_1, λ_2 with $\kappa < \lambda_1 < \lambda^* < \lambda_2$, it follows from (2.9) and Lemma 2.3(i)

that

$$\begin{aligned}
\theta^p \;&\geq\; \int_\Xi \lim_{\lambda\downarrow\lambda^*} \overline{D}(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) \quad \geq \int_\Xi \overline{D}(\lambda_2,\zeta)\boldsymbol{\nu}(d\zeta) \\
&\qquad\qquad\qquad\qquad\quad \geq \int_\Xi \underline{D}(\lambda_2,\zeta)\boldsymbol{\nu}(d\zeta) \geq \int_\Xi \mathsf{d}^p(\underline{T}(\lambda_2,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) - \delta \\
\theta^p \;&\leq \int_\Xi \lim_{\lambda\uparrow\lambda^*} \underline{D}(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) \quad \leq \int_\Xi \underline{D}(\lambda_1,\zeta)\boldsymbol{\nu}(d\zeta) \\
&\qquad\qquad\qquad\qquad\quad \leq \int_\Xi \overline{D}(\lambda_1,\zeta)\boldsymbol{\nu}(d\zeta) \leq \int_\Xi \mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + \delta
\end{aligned} \tag{2.10}$$

Based on (2.10), we now construct a feasible primal solution. Note that there is a $q^\varepsilon_\delta(\lambda_1,\lambda_2) \in [0,1]$ such that

$$q^\varepsilon_\delta(\lambda_1,\lambda_2)\left[\int_\Xi \mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta)\boldsymbol{\nu}(d\zeta)+\delta\right]+\big(1-q^\varepsilon_\delta(\lambda_1,\lambda_2)\big)\left[\int_\Xi \mathsf{d}^p(\underline{T}_{\lambda_2}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta)-\delta\right] = \theta^p. \tag{2.11}$$

Let $q^\delta := \frac{\theta^p}{\theta^p+\max\left\{0,\left[1-2q^\varepsilon_\delta(\lambda_1,\lambda_2)\right]\right\}\delta}$. Define a distribution $\boldsymbol{\mu}^\varepsilon_\delta(\lambda_1,\lambda_2)$ by

$$\boldsymbol{\mu}^\varepsilon_\delta(\lambda_1,\lambda_2) \;:=\; q^\delta q^\varepsilon_\delta(\lambda_1,\lambda_2)\overline{T}(\lambda_1,\cdot)_\#\boldsymbol{\nu} + q^\delta\big(1-q^\varepsilon_\delta(\lambda_1,\lambda_2)\big)\underline{T}(\lambda_2,\cdot)_\#\boldsymbol{\nu} + (1-q^\delta)\boldsymbol{\nu}.$$

Then $\boldsymbol{\mu}^\varepsilon_\delta(\lambda_1,\lambda_2)$ is primal feasible, because

$$\begin{aligned}
&\mathcal{W}^p_p(\boldsymbol{\mu}^\varepsilon_\delta(\lambda_1,\lambda_2),\boldsymbol{\nu}) \\
\leq\;& q^\delta q^\varepsilon_\delta(\lambda_1,\lambda_2)\int_\Xi \mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + q^\delta\big(1-q^\varepsilon_\delta(\lambda_1,\lambda_2)\big)\int_\Xi \mathsf{d}^p(\underline{T}(\lambda_2,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \\
=\;& q^\delta\Big(\theta^p + \big[1-2q^\varepsilon_\delta(\lambda_1,\lambda_2)\big]\delta\Big) \;\leq\; \theta^p
\end{aligned}$$

Furthermore, recall that

$$\begin{aligned}
\lambda_1\mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta) - \Phi(\lambda_1,\zeta) - \varepsilon \;&\leq\; \ell(\overline{T}(\lambda_1,\zeta)) \;\leq\; \lambda_1\mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta) - \Phi(\lambda_1,\zeta) \\
\lambda_2\mathsf{d}^p(\underline{T}(\lambda_2,\zeta),\zeta) - \Phi(\lambda_2,\zeta) - \varepsilon \;&\leq\; \ell(\underline{T}(\lambda_2,\zeta)) \;\leq\; \lambda_2\mathsf{d}^p(\underline{T}(\lambda_2,\zeta),\zeta) - \Phi(\lambda_2,\zeta)
\end{aligned}$$

for all $\zeta \in B$. This, together with (2.11), implies that

$$\begin{aligned}
v_P &\geq \int_\Xi \ell(\zeta)\boldsymbol{\mu}_\delta^\varepsilon(\lambda_1,\lambda_2)(d\zeta) \\
&= q^\delta q_\delta^\varepsilon(\lambda_1,\lambda_2)\int_\Xi \ell(\overline{T}(\lambda_1,\zeta))\boldsymbol{\nu}(d\zeta) \\
&\quad + q^\delta\big(1-q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\int_\Xi \ell(\underline{T}(\lambda_2,\zeta))\boldsymbol{\nu}(d\zeta) + (1-q^\delta)\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) \\
&\geq q^\delta q_\delta^\varepsilon(\lambda_1,\lambda_2)\int_\Xi \big[\lambda_1 \mathsf{d}^p(\overline{T}(\lambda_1,\zeta),\zeta) - \Phi(\lambda_1,\zeta) - \varepsilon\big]\boldsymbol{\nu}(d\zeta) + (1-q^\delta)\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) \\
&\quad + q^\delta\big(1-q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\int_\Xi \big[\lambda_2 \mathsf{d}^p(\underline{T}(\lambda_2,\zeta),\zeta) - \Phi(\lambda_2,\zeta) - \varepsilon\big]\boldsymbol{\nu}(d\zeta) \\
&\geq q^\delta\lambda_1\Big[\theta^p + \big(1-2q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\delta\Big] - q^\delta q_\delta^\varepsilon(\lambda_1,\lambda_2)\int_\Xi \Phi(\lambda_1,\zeta)\boldsymbol{\nu}(d\zeta) \\
&\quad - q^\delta\big(1-q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\int_\Xi \Phi(\lambda_2,\zeta)\boldsymbol{\nu}(d\zeta) - q^\delta\varepsilon + (1-q^\delta)\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta).
\end{aligned}$$

Recall that $\Phi(\lambda,\zeta) \leq -\ell(\zeta)$ for all $\zeta \in \Xi$. Also, consider any $\lambda_0 \in (\kappa,\lambda_1)$. Recall that $\Phi(\cdot,\zeta)$ is nondecreasing, and thus $|\Phi(\lambda,\zeta)| \leq |\ell(\zeta)| + |\Phi(\lambda_0,\zeta)|$ for all $\lambda \geq \lambda_0$ and all $\zeta \in \Xi$. Also, it follows from $\lambda_0 > \kappa$ that $\int_\Xi \Phi(\lambda_0,\zeta)\boldsymbol{\nu}(d\zeta) > -\infty$. Hence it follows from the dominated convergence theorem that $\lim_{\lambda_1\uparrow\lambda^*}\int_\Xi \Phi(\lambda_1,\zeta)\boldsymbol{\nu}(d\zeta) = \int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)$ and $\lim_{\lambda_2\downarrow\lambda^*}\int_\Xi \Phi(\lambda_2,\zeta)\boldsymbol{\nu}(d\zeta) = \int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)$. Thus

$$\begin{aligned}
&\lim_{\lambda_1\uparrow\lambda^*,\lambda_2\downarrow\lambda^*}\Big\{q_\delta^\varepsilon(\lambda_1,\lambda_2)\int_\Xi \Phi(\lambda_1,\zeta)\boldsymbol{\nu}(d\zeta) + \big(1-q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\int_\Xi \Phi(\lambda_2,\zeta)\boldsymbol{\nu}(d\zeta)\Big\} \\
&= \int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)
\end{aligned}$$

and hence

$$\begin{aligned}
v_P \geq q^\delta\lambda^*\Big[\theta^p + \limsup_{\lambda_1\uparrow\lambda^*,\lambda_2\downarrow\lambda^*}\big(1-2q_\delta^\varepsilon(\lambda_1,\lambda_2)\big)\delta\Big] \\
- q^\delta\int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) - q^\delta\varepsilon + (1-q^\delta)\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta).
\end{aligned} \tag{2.12}$$

Next, note that $1 \geq q^\delta := \frac{\theta^p}{\theta^p + \max\left\{0,\left[1-2q_\delta^\varepsilon(\lambda_1,\lambda_2)\right]\right\}\delta} \geq \frac{\theta^p}{\theta^p+\delta}$, and thus $q^\delta \to 1$ as $\delta \to 0$.

Thus taking the limit as $\delta \to 0$ in (2.12), it follows that

$$v_P \;\geq\; \lambda^*\theta^p - \int_\Xi \Phi(\lambda^*, \zeta)\boldsymbol{\nu}(d\zeta) - \varepsilon.$$

Since $\varepsilon > 0$ can be arbitrarily small, it follows that $v_P \geq \lambda^*\theta^p - \int_\Xi \Phi(\lambda^*, \zeta)\boldsymbol{\nu}(d\zeta) = v_D$.

- Case 2: κ is the unique minimizer of h.

Then h is increasing and convex on $[\kappa, \infty)$. For any $\lambda > \kappa$, it follows from h being increasing that

$$\int_\Xi \left[\Phi(\lambda, \zeta) - \Phi(\kappa, \zeta)\right] \boldsymbol{\nu}(d\zeta) \;<\; (\lambda - \kappa)\theta^p. \tag{2.13}$$

Consider any $\varepsilon \in \left(0, (\lambda - \kappa)\theta^p - \int_\Xi \left[\Phi(\lambda, \zeta) - \Phi(\kappa, \zeta)\right] \boldsymbol{\nu}(d\zeta)\right)$. It follows from Lemma 2.4 that there exists a $\boldsymbol{\nu}$-measurable map $\underline{T}_\varepsilon(\lambda, \cdot) : \Xi \to \Xi$ such that $\lambda \mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta) - \ell(\underline{T}_\varepsilon(\lambda, \zeta)) \leq \Phi(\lambda, \zeta) + \varepsilon$ for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. Also, note that $\Phi(\kappa, \zeta) \leq \kappa \mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta) - \ell(\underline{T}_\varepsilon(\lambda, \zeta))$. Thus,

$$\Phi(\lambda, \zeta) - \Phi(\kappa, \zeta) \;\geq\; (\lambda - \kappa)\mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta) - \varepsilon$$

for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. Combining this with (2.13) yields that

$$\begin{aligned}
(\lambda - \kappa) \int_\Xi \mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta)\boldsymbol{\nu}(d\zeta) \;&\leq\; \int_\Xi \left[\Phi(\lambda, \zeta) - \Phi(\kappa, \zeta)\right] \boldsymbol{\nu}(d\zeta) + \varepsilon \;<\; (\lambda - \kappa)\theta^p \\
\Rightarrow \quad \int_\Xi \mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta)\boldsymbol{\nu}(d\zeta) \;&<\; \theta^p
\end{aligned}$$

Hence, the distribution $\underline{T}_\varepsilon(\lambda, \cdot)_\#\boldsymbol{\nu}$ is primal feasible.

Next, we separately consider the cases $\kappa = 0$ and $\kappa > 0$. If $\kappa = 0$, then

$$\begin{aligned}
v_P \;\geq\; \int_\Xi \ell(\xi)\underline{T}_\varepsilon(\lambda, \cdot)_\#\boldsymbol{\nu}(d\xi) \;&\geq\; \int_\Xi [\lambda \mathsf{d}^p(\underline{T}_\varepsilon(\lambda, \zeta), \zeta) - \Phi(\lambda, \zeta) - \varepsilon]\boldsymbol{\nu}(d\zeta) \\
&\geq\; -\int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) - \varepsilon.
\end{aligned}$$

Since ε can be chosen arbitrarily small, it follows that $v_P \geq -\int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)$ for all $\lambda > \kappa = 0$. Therefore

$$v_P \geq \lim_{\lambda\downarrow 0}\left\{-\int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)\right\} = \lim_{\lambda\downarrow 0}\left\{\lambda\theta^p - \int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)\right\} = \inf_{\lambda\geq 0} h(\lambda) = v_D.$$

Otherwise, if $\kappa > 0$, then consider any $\kappa' \in (0,\kappa)$. First, note that $\zeta \in \{\xi \in \Xi : \ell(\xi) - \ell(\zeta) \geq \kappa' \mathsf{d}^p(\xi,\zeta)\}$, and hence $\{\xi \in \Xi : \ell(\xi) - \ell(\zeta) \geq \kappa' \mathsf{d}^p(\xi,\zeta)\} \neq \varnothing$ for all $\zeta \in \Xi$. Let

$$\hat{D}(\kappa',\zeta) \;:=\; \sup_{\xi\in\Xi}\{\mathsf{d}^p(\xi,\zeta) \;:\; \ell(\xi) - \ell(\zeta) \geq \kappa' \mathsf{d}^p(\xi,\zeta)\}.$$

Next we show that $\int_\Xi \hat{D}(\kappa',\zeta)\boldsymbol{\nu}(d\zeta) = \infty$. Note that $\ell(\xi) \leq \lambda_0 \mathsf{d}^p(\xi,\zeta) - \Phi(\lambda_0,\zeta)$ for all $\xi \in \Xi$. Thus

$$\begin{aligned}
\int_\Xi \Phi(\kappa',\zeta)\boldsymbol{\nu}(d\zeta) &= \int_\Xi \inf_{\xi\in\Xi}\Big\{\kappa'\mathsf{d}^p(\xi,\zeta) - \ell(\xi) \;:\; \ell(\xi) - \ell(\zeta) \geq \kappa'\mathsf{d}^p(\xi,\zeta)\Big\}\boldsymbol{\nu}(d\zeta) \\
&\geq \int_\Xi \inf_{\xi\in\Xi}\Big\{ -\ell(\xi) \;:\; \ell(\xi) - \ell(\zeta) \geq \kappa'\mathsf{d}^p(\xi,\zeta)\Big\}\boldsymbol{\nu}(d\zeta) \\
&\geq \int_\Xi \inf_{\xi\in\Xi}\Big\{ -\lambda_0\mathsf{d}^p(\xi,\zeta) + \Phi(\lambda_0,\zeta) \;:\; \ell(\xi) - \ell(\zeta) \geq \kappa'\mathsf{d}^p(\xi,\zeta)\Big\}\boldsymbol{\nu}(d\zeta) \\
&= -\lambda_0 \int_\Xi \hat{D}(\kappa',\zeta)\boldsymbol{\nu}(d\zeta) + \int_\Xi \Phi(\lambda_0,\zeta)\boldsymbol{\nu}(d\zeta).
\end{aligned}$$

It follows from the definition of κ that $\int_\Xi \Phi(\kappa',\zeta)\boldsymbol{\nu}(d\zeta) = -\infty$, and therefore $\int_\Xi \hat{D}(\kappa',\zeta)\boldsymbol{\nu}(d\zeta) = \infty$. It follows that for any $R > 0$, there exists $M \in L^1(\boldsymbol{\nu})$ such that $\int_\Xi M(\zeta)\boldsymbol{\nu}(d\zeta) > R$ and

$$\{\xi \in \Xi \;:\; \ell(\xi) - \ell(\zeta) \geq \kappa'\mathsf{d}^p(\xi,\zeta),\; \mathsf{d}^p(\xi,\zeta) \geq M(\zeta)\} \;\neq\; \varnothing$$

for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. Then it follows from Lemma 2.4(v) that for any $\kappa' \in (0,\kappa)$ and $R > \theta^p$, there exists a $\boldsymbol{\nu}$-measurable mapping $T^R_{\kappa'} : \Xi \to \Xi$ such that

$$\ell\big(T^R_{\kappa'}(\zeta)\big) - \ell(\zeta) \;\geq\; \kappa'\mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)$$

for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$, and

$$\int_{\Xi} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \;\geq\; \int_{\Xi} M(\zeta)\boldsymbol{\nu}(d\zeta) \;>\; R.$$

If $\int_{\Xi} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) = \infty$, let $F_r := \{\zeta \in \Xi : \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta) \leq r\}$. Note that $\lim_{r\to\infty} F_r = \Xi$, and thus there exists a $\bar{r} > R$ such that $R < \int_{F_{\bar{r}}} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \leq \bar{r} < \infty$. Then, for all $\zeta \in F_{\bar{r}}$, let $\overline{T}^R_{\kappa'}(\zeta) := T^R_{\kappa'}(\zeta)$, and for all $\zeta \in \Xi \setminus F_{\bar{r}}$, let $\overline{T}^R_{\kappa'}(\zeta) := \zeta$. Note that $\overline{T}^R_{\kappa'}$ is $\boldsymbol{\nu}$-measurable, and that $\int_{\Xi} \mathsf{d}^p(\overline{T}^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) = \int_{F_{\bar{r}}} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \in (R,\infty)$. Otherwise, if $\int_{\Xi} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) < \infty$, then let $F_{\bar{r}} := \Xi$ and $\overline{T}^R_{\kappa'}(\zeta) := T^R_{\kappa'}(\zeta)$ for all $\zeta \in \Xi$. Let $q^R_{\varepsilon,\kappa'} \in (0,1)$ be such that

$$q^R_{\varepsilon,\kappa'} \int_{\Xi} \mathsf{d}^p(\underline{T}_{\varepsilon}(\lambda,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + (1-q^R_{\varepsilon,\kappa'}) \int_{\Xi} \mathsf{d}^p(\overline{T}^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \theta^p.$$

Next we construct a primal feasible solution

$$\boldsymbol{\mu}^R_{\varepsilon,\kappa'}(\lambda) \;:=\; q^R_{\varepsilon,\kappa'}\underline{T}_{\varepsilon}(\lambda,\cdot)_{\#}\boldsymbol{\nu} + (1-q^R_{\varepsilon,\kappa'})(\overline{T}^R_{\kappa'})_{\#}\boldsymbol{\nu}$$

By construction, $\boldsymbol{\mu}^R_{\varepsilon,\kappa'}(\lambda)$ is primal feasible. Moreover,

$$\begin{aligned}
v_P \;&\geq\; \int_{\Xi} \ell(\xi)\boldsymbol{\mu}^R_{\varepsilon,\kappa'}(\lambda)(d\xi) \\
&=\; q^R_{\varepsilon,\kappa'} \int_{\Xi} \ell(\underline{T}_{\varepsilon}(\lambda,\zeta))\boldsymbol{\nu}(d\zeta) + (1-q^R_{\varepsilon,\kappa'}) \int_{\Xi} \ell(\overline{T}^R_{\kappa'}(\zeta))\boldsymbol{\nu}(d\zeta) \\
&\geq\; q^R_{\varepsilon,\kappa'} \int_{\Xi} \big[\lambda \mathsf{d}^p(\underline{T}_{\varepsilon}(\lambda,\zeta),\zeta) - \Phi(\lambda,\zeta) - \varepsilon\big]\boldsymbol{\nu}(d\zeta) \\
&\quad +(1-q^R_{\varepsilon,\kappa'}) \int_{F_{\bar{r}}} \big[\kappa' \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta) + \ell(\zeta)\big]\boldsymbol{\nu}(d\zeta) + (1-q^R_{\varepsilon,\kappa'}) \int_{\Xi\setminus F_{\bar{r}}} \ell(\zeta)\boldsymbol{\nu}(d\zeta) \\
&\geq\; \kappa' \left(q^R_{\varepsilon,\kappa'} \int_{\Xi} \mathsf{d}^p(\underline{T}_{\varepsilon}(\lambda,\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + (1-q^R_{\varepsilon,\kappa'}) \int_{F_{\bar{r}}} \mathsf{d}^p(T^R_{\kappa'}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \right) \\
&\quad -q^R_{\varepsilon,\kappa'} \int_{\Xi} \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) - q^R_{\varepsilon,\kappa'}\varepsilon + (1-q^R_{\varepsilon,\kappa'}) \int_{\Xi} \ell(\zeta)\boldsymbol{\nu}(d\zeta) \\
&=\; \kappa'\theta^p - q^R_{\varepsilon,\kappa'} \int_{\Xi} \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) - q^R_{\varepsilon,\kappa'}\varepsilon + (1-q^R_{\varepsilon,\kappa'}) \int_{\Xi} \ell(\zeta)\boldsymbol{\nu}(d\zeta).
\end{aligned}$$

For any fixed $\lambda > \kappa$, $\varepsilon > 0$, and $\kappa' \in (0, \kappa)$, it holds that $q^R_{\varepsilon,\kappa'}$ can be chosen arbitrarily close to 1 by choosing sufficiently large R. Hence $v_P \geq \kappa'\theta^p - \int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) - \varepsilon$ for all $\lambda > \kappa$, $\varepsilon > 0$, and $\kappa' \in (0, \kappa)$. Since $\varepsilon > 0$ can be chosen arbitrarily small and κ' can be chosen arbitrarily close to κ, it follows that $v_P \geq \kappa\theta^p - \int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta)$ for all $\lambda > \kappa$. Therefore $v_P \geq \lim_{\lambda \downarrow \kappa} \left\{\lambda\theta^p - (\lambda - \kappa)\theta^p - \int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta)\right\} \geq \inf_{\lambda > \kappa}\{\lambda\theta^p - \int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta)\} = v_D$. □

Remark 2.2. All the above results and proofs in Section 2.4.1 (except for Lemma 2.2 continue to hold if we replace the transportation cost $\mathsf{d}^p(\cdot, \cdot)$ with any measurable, nonnegative cost function $c(\cdot, \cdot)$ that satisfies $c(\xi, \zeta) = 0$ if $\xi = \zeta$.

Next, we investigate existence conditions for worst-case distributions and their structure. In the rest of this section, we assume that ℓ is upper-semicontinuous, and every bounded subset in (Ξ, d) is totally bounded, which is satisfied by, for example, any finite-dimensional normed space. Under this assumption, if $\lambda > \kappa$ and $\zeta \in B$, then Lemma 2.3(ii) and the upper semi-continuity of ℓ imply that the set $\arg\min_{\xi \in \Xi}\{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\}$ is nonempty, and that $\min / \max_{\xi \in \Xi}\{\mathsf{d}^p(\xi, \zeta) : \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) = \Phi(\lambda, \zeta)\}$ can be attained. If $\lambda = \kappa$ and $\boldsymbol{\nu}\big(\{\zeta \in \Xi : \arg\min_{\xi \in \Xi}\{\kappa \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\} = \varnothing\}\big) = 0$, then the upper semi-continuity of ℓ imply that $\min_{\xi \in \Xi}\{\mathsf{d}^p(\xi, \zeta) : \kappa \mathsf{d}^p(\xi, \zeta) - \ell(\xi) = \Phi(\kappa, \zeta)\}$ can be attained for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$, but $\sup_{\xi \in \Xi}\{\mathsf{d}^p(\xi, \zeta) : \kappa \mathsf{d}^p(\xi, \zeta) - \ell(\xi) = \Phi(\kappa, \zeta)\}$ can be infinite. Thus, if (i) $\lambda > \kappa$, or (ii) $\lambda = \kappa$ and $\boldsymbol{\nu}\big(\{\zeta \in \Xi : \arg\min_{\xi \in \Xi}\{\kappa \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\} = \varnothing\}\big) = 0$, then the quantities $\underline{D}_0(\lambda, \zeta)$ and $\overline{D}_0(\lambda, \zeta)$ in (2.4) are well-defined for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$ (where $\overline{D}_0(\lambda, \zeta)$ can be infinite if $\lambda = \kappa$).

Corollary 2.1 (Worst-case distribution)**.** *Consider any $p \in [1, \infty)$, $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, $\theta > 0$, and $\ell \in L^1(\boldsymbol{\nu})$ such that $\kappa < \infty$. Assume that ℓ is upper-semicontinuous, and that bounded subsets of (Ξ, d) are totally bounded. Then the following holds:*

(1) [Existence condition] A worst-case distribution exists if and only if any of the following conditions hold:

(i) There exists a dual minimizer $\lambda^* > \kappa$.

(ii) $\lambda^* = \kappa > 0$ is the unique dual minimizer, $\boldsymbol{\nu}\big(\{\zeta \in \Xi : \arg\min_{\xi\in\Xi}\{\kappa \mathsf{d}^p(\xi,\zeta) - \ell(\xi)\} = \varnothing\}\big) = 0$, and

$$\int_\Xi \underline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta) \;\; \leq \;\; \theta^p \;\; \leq \;\; \int_\Xi \overline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta).$$

(iii) $\lambda^* = \kappa = 0$ is the unique dual minimizer, $\arg\max_{\xi\in\Xi}\{\ell(\xi)\}$ is nonempty, and

$$\int_\Xi \underline{D}_0(0,\zeta)\boldsymbol{\nu}(d\zeta) \;\; \leq \;\; \theta^p.$$

(2) If $\boldsymbol{\nu}\big(\zeta \in \Xi : -\ell(\zeta) > \inf_{\xi\in\Xi}\{\kappa \mathsf{d}^p(\xi,\zeta) - \ell(\xi)\}\big) = 0$, then $\lambda^* = \kappa$ for any $\theta > 0$. Otherwise, there is $\theta_0 > 0$ such that $\lambda^* > \kappa$ for any $\theta < \theta_0$.

(3) [Structure] Whenever a worst-case distribution exists, there exists a worst-case distribution $\boldsymbol{\mu}^*$ which can be represented as a convex combination of two distributions $\overline{T}^*_\# \boldsymbol{\nu}$ and $\underline{T}^*_\# \boldsymbol{\nu}$, each of which is a perturbation of $\boldsymbol{\nu}$, as follows:

$$\boldsymbol{\mu}^* \;\; = \;\; p^* \overline{T}^*_\# \boldsymbol{\nu} + (1-p^*)\underline{T}^*_\# \boldsymbol{\nu},$$

where $p^* \in [0,1]$, and $\overline{T}^*, \underline{T}^* : \Xi \to \Xi$ satisfy

$$\boldsymbol{\nu}\big(\zeta \in \Xi : \overline{T}^*(\zeta), \underline{T}^*(\zeta) \notin \arg\min_{\xi\in\Xi}\{\lambda^* \mathsf{d}^p(\xi,\zeta) - \ell(\xi)\}\big) \;\; = \;\; 0 \qquad (2.14)$$

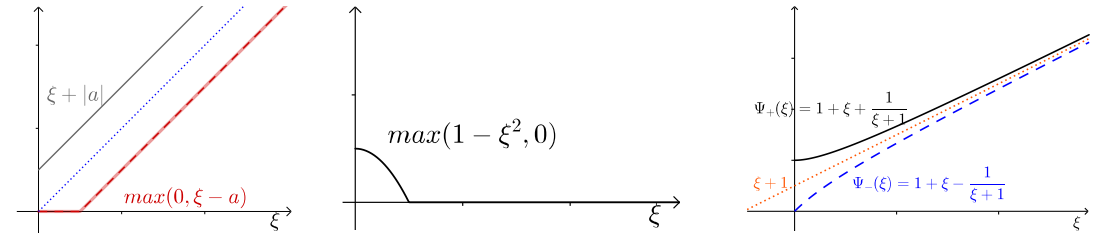
(4) If Ξ is a convex subset of a Banach space and ℓ is concave, then

$$v_P = v_D = \sup_{T:\Xi\to\Xi}\Big\{\mathbb{E}_{T_\#\boldsymbol{\nu}}[\ell(\xi)] : \mathcal{W}_p(T_\#\nu,\nu) \leq \theta\Big\}.$$

Moreover, whenenver the worst-case distribution exists, there exists $T^* : \Xi \to \Xi$ such that $T^*_\# \boldsymbol{\nu}$ is primal optimal.

Remark 2.3. Compared with Corollary 4.7 in [24], Corollary 2.1(1) provides a complete description of the necessary and sufficient conditions for the existence of a worst-case distribution. Note that Example 1 in [24] corresponds to $\lambda^* = \kappa = 1$ and $p = 1$.

Example 2.4. We present several examples that correspond to different cases in Corollary 2.1(1). In all these examples, $\Xi = [0, \infty)$, $\mathsf{d}(\xi, \zeta) = |\xi - \zeta|$ for all $\xi, \zeta \in \Xi$, $p = 1$, $\theta > 0$, and $\boldsymbol{\nu} = \boldsymbol{\delta}_0$.



(a) $\ell_a(\xi) = \max\{0, \xi - a\}$ (b) $\ell(\xi) = \max\{0, 1 - \xi^2\}$ (c) $\ell_\pm(\xi) = 1 + \xi \pm \frac{1}{\xi+1}$

Figure 2.2: Examples for existence and non-existence of the worst-case distribution

1. $\ell_a(\xi) := \max\{0, \xi - a\}$ for some $a \in \mathbb{R}$. It follows that $\lambda^* = \kappa = 1$.

   – If $a \leq 0$, then $\arg\min_{\xi \in \Xi}\{\mathsf{d}^p(\xi, 0) - \ell_a(\xi)\} = [0, \infty)$, hence $\underline{D}_0(\kappa, \zeta) = 0$ and $\overline{D}_0(\kappa, \zeta) = \infty$. Thus condition (ii) is satisfied. One of the worst-case distributions is $\boldsymbol{\mu}^* = \boldsymbol{\delta}_\theta$ with $v_P = v_D = \theta - a$.

   – If $a > 0$, then $\arg\min_{\xi \in \Xi}\{\mathsf{d}^p(\xi, 0) - \ell_a(\xi)\} = \{0\}$, hence $\underline{D}_0(\kappa, \zeta) = \overline{D}_0(\kappa, \zeta) = 0 < \theta$. Thus condition (ii) is violated. There is no worst-case distribution, but the objective value of $\boldsymbol{\mu}_\varepsilon = (1 - \varepsilon)\boldsymbol{\delta}_0 + \varepsilon\boldsymbol{\delta}_{\theta/\varepsilon}$ converges to $v_P = v_D = \theta$ as $\varepsilon \to 0$.

2. $\ell(\xi) = \max\{0, 1 - \xi^2\}$. It follows that $\lambda^* = \kappa = 0$, and $\arg\max_{\xi \in \Xi} \ell(\xi) = \{0\}$. Thus condition (iii) is satisfied, and the worst-case distribution is $\boldsymbol{\mu}^* = \boldsymbol{\delta}_0 = \boldsymbol{\nu}$.

3. $\ell_\pm(\xi) = 1 + \xi \pm \frac{1}{\xi+1}$. It follows that $\kappa = 1$. Note that $\ell'_\pm(\xi) = 1 \mp \frac{1}{(\xi+1)^2}$.

   – Note that $\ell'_+(\xi) < \kappa = 1$ on Ξ. Also, ℓ_+ satisfies the condition in (2), thus for all $\theta > 0$ it holds that $\lambda^*_+ = \kappa = 1$ and $\arg\min_{\xi \in \Xi}\{\lambda^*_+ \mathsf{d}^p(\xi, 0) - \ell_+(\xi)\} = \{0\}$. There

is no worst-case distribution, but the objective value of $\boldsymbol{\mu}_\varepsilon = (1-\varepsilon)\boldsymbol{\delta}_0 + \varepsilon\boldsymbol{\delta}_{\theta/\varepsilon}$ converges to $v_P = v_D = 2+\theta$ as $\varepsilon \to 0$.

– Note that $\ell'_-(\xi) > \kappa = 1$ on Ξ. Also, $\arg\min_{\lambda\geq 0}\{\lambda\theta - \inf_{\xi\in\Xi}\{\lambda\xi - (1+\xi - \frac{1}{\xi+1})\}\} = \arg\min_{\lambda\geq 1}\{\lambda(\theta+1) - 2\sqrt{\lambda-1}\} = \{1+\frac{1}{(\theta+1)^2}\}$. Thus $\lambda^*_- > 1 = \kappa$.

2.4.2 Finite-Supported Nominal Distribution

In this section, we restrict attention to the case in which the nominal distribution $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\widehat{\xi}^i}$ for some $\widehat{\xi}^i \in \Xi$, $i = 1,\ldots,n$. This occurs, for example, in a data-driven setting in which the decision maker collects n observations that constitute an empirical distribution.

Corollary 2.2 (Data-Driven DRSO)**.** *Consider any $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\widehat{\xi}^i}$. Let $p \in [1,\infty)$ and $\theta > 0$. Then the following hold:*

(i) [Strong duality] The primal problem (Primal) has a strong dual problem

$$v_P = v_D = \inf_{\lambda\geq 0}\left\{\lambda\theta^p + \frac{1}{n}\sum_{i=1}^n \sup_{\xi\in\Xi}\left[\ell(\xi) - \lambda \mathsf{d}^p(\xi,\widehat{\xi}^i)\right]\right\}. \tag{2.15}$$

(ii) [Structure of the worst-case distribution] Whenever a worst-case distribution exists, there exists one which is supported on at most $n+1$ points and has the form

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i\neq i_0}\boldsymbol{\delta}_{\xi^i_*} + \frac{p_0}{n}\boldsymbol{\delta}_{\underline{\xi}^{i_0}_*} + \frac{1-p_0}{n}\boldsymbol{\delta}_{\overline{\xi}^{i_0}_*} \tag{2.16}$$

where $i_0 \in \{1,\ldots,n\}$, $p_0 \in [0,1]$, $\underline{\xi}^{i_0}_*, \overline{\xi}^{i_0}_* \in \arg\min_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\widehat{\xi}^{i_0}) - \ell(\xi)\}$, and $\xi^i_* \in \arg\min_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\widehat{\xi}^i) - \ell(\xi)\}$ for all $i \neq i_0$.

(iii) [Robust-program approximation] Suppose that there exists $\zeta^0 \in \Xi$, $L, M \geq 0$ such that $|\ell(\xi) - \ell(\zeta^0)| < L\mathsf{d}^p(\xi,\zeta^0) + M$ for all $\xi \in \Xi$. Let K be any positive integer

and consider the robust optimization problem

$$v_K \;:=\; \sup_{(\xi^{ik})_{i,k}\in\mathfrak{M}_K} \frac{1}{Kn}\sum_{i=1}^{n}\sum_{k=1}^{K}\ell(\xi^{ik}),$$

with uncertainty set

$$\mathfrak{M}_K \;:=\; \left\{(\xi^{ik})_{i,k} \;:\; \frac{1}{Kn}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathsf{d}^p(\xi^{ik},\widehat{\xi}^i)\le\theta^p,\; \xi^{ik}\in\Xi\;\forall\, i,k\right\}.$$

If $\lambda^* > \kappa$, then there exists $L', M' > 0$, such that

$$v_K \;\le\; \sup_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\mu}}[\ell(\xi)] \;\le\; v_K + \frac{M' + L'D}{Kn},$$

where D is a constant independent of K. In addition, if Ξ is convex and ℓ is concave, then $v_1 = v_P = v_D$.

Statement (ii) shows that the worst-case distribution $\boldsymbol{\mu}^*$ is a *perturbation* of $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}_{\widehat{\xi}^i}$, where $n-1$ out of the n points, $\{\widehat{\xi}^i\}_{i\neq i_0}$, are perturbed with full mass $1/N$ to a maximizer ξ_*^i respectively, while at most one point $\widehat{\xi}^{i_0}$ is split and perturbed to two maximizers $\underline{\xi}_*^{i_0}$ and $\overline{\xi}_*^{i_0}$. (If the set of maximizers is a singleton, then there is no need to split). Using this structure, we obtain statement (iii), which suggests that the primal problem can be approximated by a robust program with uncertainty set $\mathfrak{M}_K$, which is a subset of $\mathfrak{M}$ that contains all distributions supported on Kn points with equal probability $\frac{1}{Kn}$. Particularly, when ℓ is concave, such approximation is exact; and when ℓ is Lipschitz and $p = 1$, then v_1 is an $O(1/n)$-approximation of $v_P = v_D$.

Remark 2.4. The results in Corollary 2.2 hold for arbitrary metric space Ξ. In fact, the Polish space assumption on Ξ is only used for the measurability results in Lemma 2.4, which becomes trivial in finite-supported case.

Remark 2.5. Under compactness assumption on Ξ, [22] pointed out that to solve (Primal),

it suffices to consider the set of extreme points of the Wasserstein ball $\mathfrak{M}$ contains distributions that are supported on at most $n+3$ points. Later in [64], this result was improved to $n+2$ for Polish space or Borel subsets of Polish space. Statement (ii) further strength these results — for arbitrary metric space (see Remark 2.4 above), it suffices to consider distributions that are supported on at most $n+1$ points, and such bound is tight as shown by Example 2.7 below. Moreover, the weight of the extreme distribution do not change much as compared to the nominal distribution. As can be immediately seen from the proof, the result of statement (ii) can be generalized as following. Suppose $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^{n} \nu_i \boldsymbol{\delta}_{\widehat{\xi}^i}$, then whenever the worst-case distribution exists, there exists one of the form

$$\sum_{i \neq i_0} \nu_i \boldsymbol{\delta}_{\xi_*^i} + p_0 \nu_{i_0} \boldsymbol{\delta}_{\underline{\xi}_*^{i_0}} + (1-p_0)\nu_{i_0} \boldsymbol{\delta}_{\overline{\xi}_*^{i_0}}.$$

Remark 2.6 (Total Variation metric)**.** By choosing the discrete metric $\mathsf{d}(\xi,\zeta) = \mathbb{1}_{\{\xi \neq \zeta\}}$ on Ξ, the Wasserstein distance is equal to Total Variation distance [74], which can be used for the situation where the distance of perturbation does not matter and provides a rather conservative decision. In this case, suppose θ is chosen such that $n\theta$ is an integer, then there is no fractional point in (2.16) and the problem is reduced to the robust program with uncertainty set $\mathfrak{M}_1$, whether Ξ (ℓ) is convex (concave) or not.

Proof. Proof of Corollary 2.2.

(i) It follows directly from the proof of Theorem 2.1 and Proposition 2.2.

(ii) By Corollary 2.1(3), whenever the worst-case distribution exists, there is one supported on at most $2n$ points and has the form

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i=1}^{n} p^i \boldsymbol{\delta}_{\underline{\xi}_*^i} + (1-p^i)\boldsymbol{\delta}_{\overline{\xi}_*^i}, \tag{2.17}$$

where $p^i \in [0,1]$, and $\underline{\xi}_*^i, \overline{\xi}_*^i \in \arg\min_{\xi \in \Xi}\{\lambda^* \mathsf{d}^p(\xi, \widehat{\xi}^i) - \ell(\xi)\}$. (In fact, Corollary 2.1(3) proves a stronger statement that there exists a worst-case distribution such that all p^i are

equal, but here we allow them to vary in order to obtain a worst-case distribution with a different form.) Given $\underline{\xi}_*^i, \overline{\xi}_*^i$ for all i, the problem

$$\max_{0\le p^i\le 1} \Big\{ \frac{1}{n}\sum_{i=1}^n p^i \ell(\underline{\xi}_*^i) + (1-p^i)\ell(\overline{\xi}_*^i) : \ \frac{1}{n}\sum_{i=1}^n p^i \mathsf{d}^p(\underline{\xi}_*^i, \widehat{\xi}^i) + (1-p^i)\mathsf{d}^p(\overline{\xi}_*^i, \widehat{\xi}^i) \le \theta^p \Big\}$$

is a linear program with n variables, one equality constraint and $2n$ inequality constraints $p_i \le 1$, $p_i \ge 1$, $i = 1, \ldots, n$. Thus according to linear programming theory, there exists an optimal solution such that among the $2n$ inequality constraints, at least $n-1$ of them hold as equality, or equivalently, at most one p_i is fractional. Therefore there exists a worst-case distribution which is supported on at most $n+1$ points, and has the form (2.16).

(iii) Note that by assumption on ℓ and Lemma 2.2 we have $\kappa \le L < \infty$. Also note that using the similar idea the above proof of (ii), the distributions $\boldsymbol{\mu}_\delta^\varepsilon(\lambda_1, \lambda_2)$, $(\overline{T}_\lambda)_\# \boldsymbol{\nu}$ and $\boldsymbol{\mu}_\delta^R(\lambda, \varepsilon)$ defined in the proof of Theorem 2.1 can be written in the form of

$$\frac{1}{n}\sum_{i=1}^n p^{i1}\boldsymbol{\delta}_{\xi^{i1}} + p^{i2}\boldsymbol{\delta}_{\xi^{i2}} + p^{i3}\boldsymbol{\delta}_{\xi^{i3}},$$

where $p_{i1} + p_{i2} + p_{i3} = 1$. Given $\{\xi^{ij} : 1 \le i \le n, 1 \le j \le 3\}$, the problem

$$\max_{0\le p^{ij}\le 1} \Big\{ \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^3 p^{ij} \ell(\underline{\xi}^{ij}) : \ \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^3 p^{ij} \mathsf{d}^p(\underline{\xi}^{ij}, \widehat{\xi}^i) \le \theta^p \Big\}$$

is a linear program with $3n$ variables, one equality constraint and $3n$ inequality constraints $p_{ij} \le 1$, $p_{ij} \ge 1$, $i = 1, \ldots, n$, $j = 1, 2, 3$. Thus according to linear programming theory, there exists an optimal solution such that among the $3n$ inequality constraints, at least $3N-1$ of them hold as equality, or equivalently, at most one p_{ij} is fractional. Hence for any ε-optimal solution $\boldsymbol{\mu}$, there exists a solution of the form

$$\boldsymbol{\mu}_\varepsilon = \frac{1}{n}\sum_{i\ne i_0} \boldsymbol{\delta}_{\xi_\varepsilon^i} + \frac{p_\varepsilon}{n}\boldsymbol{\delta}_{\underline{\xi}_\varepsilon^{i_0}} + \frac{1-p_\varepsilon}{n}\boldsymbol{\delta}_{\overline{\xi}_\varepsilon^{i_0}},$$

which yields an objective value no worse than $\boldsymbol{\mu}$. Define

$$\xi^{ik} = \begin{cases} \xi^i_\varepsilon, & \forall\, 1 \le k \le K,\ \forall\, i \ne i_0, \\ \underline{\xi}^{i_0}_\varepsilon, & \forall\, 1 \le k \le \lfloor K p_\varepsilon \rfloor,\ i = i_0, \\ \overline{\xi}^{i_0}_\varepsilon, & \forall\, \lfloor K p_\varepsilon \rfloor < k \le n,\ i = i_0. \end{cases}$$

Then $\{\xi^{ik}\}_{i,k}$ belongs to $\mathfrak{M}_K$. By Lemma 2.3(ii), for any $\lambda > \lambda_0 \in \mathrm{dom}(\Phi(\cdot, \widehat{\xi}^{i0}))$,

$$\mathsf{d}^p(\underline{\xi}^{i_0}_\varepsilon, \widehat{\xi}^{i_0}) \ \le\ -\frac{1}{\lambda - \lambda_0}\big(\ell(\widehat{\xi}^{i_0}) + \Phi(\lambda_0, \widehat{\xi}^{i_0})\big) \ =:\ D.$$

Since $\big|p_\varepsilon - \lfloor K p_\varepsilon \rfloor / K\big| < 1/K$, it follows that

$$\begin{aligned} \big|v_K - \mathbb{E}_{\boldsymbol{\mu}_\varepsilon}[\ell(\xi)]\big| &\le \frac{1}{n}\big|p_\varepsilon - \lfloor K p_\varepsilon \rfloor / K\big| \cdot \big(\ell(\underline{\xi}^{i_0}_\varepsilon) - \ell(\overline{\xi}^{i_0}_\varepsilon)\big) \\ &\le \frac{1}{Kn}\big(\ell(\underline{\xi}^{i_0}_\varepsilon) - \ell(\widehat{\xi}^{i_0})\big) \\ &\le \frac{M + L\mathsf{d}^p(\underline{\xi}^{i_0}_\varepsilon, \widehat{\xi}^{i_0})}{Kn} \\ &\le \frac{M + LD}{Kn}. \end{aligned}$$

Let $\varepsilon \to 0$ we obtain the results. □

Example 2.5 (Saddle-point Problem)**.** When $\ell(\beta, \xi)$ is convex in β and concave ξ, $p = 1$, and $\mathsf{d}(\cdot, \cdot) = ||\cdot - \cdot||_2$, Corollary 2.2(iii) shows that the DRSO (Wasserstein-DRSO) is equivalent to a convex-concave saddle point problem

$$\min_{\beta \in \mathcal{D}} \max_{(\xi^1, \ldots, \xi^n) \in Y} \frac{1}{n} \sum_{i=1}^n \ell(\beta, \xi^i),$$

with ℓ_1/ℓ_2-norm uncertainty set

$$Y = \left\{ (\xi^1, \ldots, \xi^n) \in \Xi^n : \sum_{i=1}^n ||\xi^i - \widehat{\xi}^i||_2 \le n\theta \right\}.$$

Therefore it can be solved by the Mirror-Prox algorithm (cf. [75, 76] and Appendix A.4).

Example 2.6 (Piecewise concave objective)**.** [24] proves that when $p = 1$, Ξ is a convex subset of $\mathbb{R}^K$ equipped with some norm $||\cdot||$ and $\ell(\xi) = \max_{1\le j\le J} \ell^j(\xi)$, where ℓ^j are concave, the DRSO is equivalent to a convex program. We here show that it can be obtained as a corollary from the structure of the worst-case distribution. Indeed, using concavity of ℓ^j and Corollary 2.2(i), it suffices to consider distributions of the form

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}\boldsymbol{\delta}_{\xi^{ij}},\ \ \sum_{j=1}^{J} p_{ij} = 1,$$

where for each i,

$$\mathrm{card}\{j : p_{ij} > 0\} \le 2,$$

where card represents cardinality. Relaxing the cardinality constraints yields the following problem:

$$\sup_{p_{ij}\ge 0, \xi^{ij}\in\Xi} \left\{ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}\ell(\xi^{ij}) :\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij} d(\xi^{ij}, \widehat{\xi^i}) \le \theta,\ \sum_{j=1}^{J} p_{ij} = 1, \forall i \right\}.$$

Replacing ξ^{ij} by $\widehat{\xi^i} + (\xi^{ij} - \widehat{\xi^i})/p_{ij}$, by positive homogeneity of norms and convexity-preserving property of perspective functions (cf. Section 2.3.3 in [77]), we obtain an equivalent convex program reformulation of the primal problem:

$$\sup_{\substack{p_{ij}\ge 0, \sum_j p_{ij}=1 \\ \xi^{ij}\in\mathbb{R}^K}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} p_{ij}\ell^j(\widehat{\xi^i} + \tfrac{\xi^{ij}-\widehat{\xi^i}}{p_{ij}}) :\ \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J} d(\xi^{ij}, \widehat{\xi^i}) \le \theta,\ \widehat{\xi^i} + \tfrac{\xi^{ij}-\widehat{\xi^i}}{p_{ij}} \in \Xi, \forall i, j \right\}.$$

So we recover Theorem 4.5 in [24], which was obtained therein by a separate procedure of dualizing twice the reformulation (2.15).

Example 2.7 (**Uncertainty Quantification**)**.** When $\Xi = \mathbb{R}^K$ and $\ell = -\mathbb{1}_C$, where C is an

open set, the worst-case distribution $\boldsymbol{\mu}^*$ of the problem

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\mu}}[-\mathbb{1}_C(\xi)] \;=\; \min_{\boldsymbol{\mu}\in\mathfrak{M}} \boldsymbol{\mu}(C)$$

has a clear interpretation. The worst-case distribution perturbs $\boldsymbol{\nu}$ such that the set C contains as little probability mass as possible, which can be achieved in a greedy fashion as follows. Suppose $\{\widehat{\xi}^i\}_{i=1}^n$ are sorted such that $\widehat{\xi}^1,\ldots,\widehat{\xi}^I \in C$, $\widehat{\xi}^{I+1},\ldots,\widehat{\xi}^n \notin C$ and satisfy $\mathsf{d}^p(\widehat{\xi}^1,\Xi\setminus C) \le \cdots \le \mathsf{d}^p(\widehat{\xi}^I,\Xi\setminus C)$. Then to save the total budget of perturbation $\widehat{\xi}^{I+1},\ldots,\widehat{\xi}^n$ stay at the same place, and the $\widehat{\xi}^i$ with small index has the priority to be transported to ∂C. It may happen that some point $\widehat{\xi}^{i_0}$ ($i_0 \le I$) cannot be transported to ∂C with full mass $\frac{1}{n}$, since otherwise the Wasserstein distance constraint is violated. In this case, only partial mass (with probability p_0/N) is transported and the remaining stays (see Figure 2.3). Therefore the worst-case distribution has the form

$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i=1}^{i_0-1} \boldsymbol{\delta}_{\xi_*^i} + \frac{p_0}{n}\boldsymbol{\delta}_{\widehat{\xi}^{i_0}} + \frac{1-p_0}{n}\boldsymbol{\delta}_{\xi_*^{i_0}} + \frac{1}{n}\sum_{i=i_0+1}^{n} \boldsymbol{\delta}_{\widehat{\xi}^i}.$$

In fact, the dual optimizer λ^* is such that

$$\xi_*^i = \operatorname*{arg\,min}_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\widehat{\xi}^i) + \mathbb{1}_C(\xi)\} = \operatorname*{arg\,min}_{\xi\in\partial C} \mathsf{d}^p(\xi,\widehat{\xi}^i), \;\; \forall i \le I,$$

and

$$\xi_*^{i_0} = \operatorname*{arg\,min}_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\widehat{\xi}^{i_0}) + \mathbb{1}_C(\xi)\} = \begin{cases} \{\widehat{\xi}^{i_0}\} \cup \arg\min_{\xi\in\partial C} \mathsf{d}^p(\xi,\widehat{\xi}^{i_0}), & p_0 \neq 0, \\ \arg\min_{\xi\in\partial C} \mathsf{d}^p(\xi,\widehat{\xi}^{i_0}), & p_0 = 0. \end{cases}$$

Using the similar idea as above, we can prove that the worst-case probability is continuous with respect to the boundary.

Proposition 2.3 (Continuity with respect to the boundary)**.** *Let* $\Xi = \mathbb{R}^K$, $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$, $\theta > 0$,

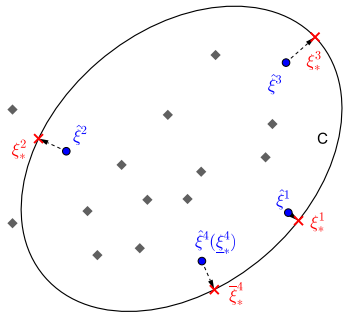


Figure 2.3: When $\ell = -\mathbb{1}_C$, the worst-case distribution perturbs the nominal distribution in a greedy fashion. The solid and diamond dots are the support of nominal distribution $\boldsymbol{\nu}$. $\widehat{\xi}^1, \widehat{\xi}^2, \widehat{\xi}^3$ are three closest interior points to ∂C and thus are transported to $\xi_*^1, \xi_*^2, \xi_*^3$ respectively. $\widehat{\xi}^4$ is the fourth closest interior point to ∂C, but cannot be transported to ∂C as full mass due to Wasserstein distance constraint, so it is split into $\overline{\xi}_*^4$ and $\underline{\xi}_*^4$.

and $\mathfrak{M} = \{\boldsymbol{\mu} \in \mathcal{P}(\Xi) \,:\, \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \theta\}$. *Then for any Borel set* $C \subset \Xi$,

$$\inf_{\boldsymbol{\mu} \in \mathfrak{M}} \boldsymbol{\mu}(C) = \min_{\boldsymbol{\mu} \in \mathfrak{M}} \boldsymbol{\mu}(\mathrm{int}(C)).$$

Now let us consider a special case when $\Xi = \{\xi^0, \ldots, \xi^{\bar{B}}\}$ for some positive integer $\bar{B}$. In this case, let n_i be the samples that are equal to ξ^i, and let $q_i = n_i/n$, $i = 0, \ldots, \bar{B}$, then the nominal distribution $\boldsymbol{\nu} = \sum_{i=1}^{\bar{B}} q_i \boldsymbol{\delta}_{\xi^i}$. Let $\boldsymbol{\nu} := (q_0, \ldots, q_{\bar{B}})^\top \in \Delta_{\bar{B}}$. The DRSO becomes

$$\min_{\beta \in \mathcal{D}} \max_{\boldsymbol{\mu} \in \Delta_{\bar{B}}} \left\{ \sum_{i=0}^{\bar{B}} p_i \ell(\beta, \xi^i) : \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \theta \right\}. \tag{2.18}$$

Corollary 2.3. *Problem (2.18) has a strong dual*

$$\min_{\beta \in \mathcal{D}, \lambda \ge 0} \left\{ \lambda \theta^p + \sum_{i=0}^{\bar{B}} q_i y_i : \; y_i \ge \ell(\beta, \xi^j) - \lambda \mathsf{d}^p(\xi^i, \xi^j), \; \forall i, j = 1, \ldots, \bar{B} \right\}. \tag{2.19}$$

For any β, *the worst-case distribution can be computed by*

$$\max_{\boldsymbol{\mu} \in \Delta_{\bar{B}}, \boldsymbol{\gamma} \in \mathbb{R}_+^{\bar{B} \times \bar{B}}} \left\{ \sum_{i=0}^{\bar{B}} p_i \ell(\beta, \xi^i) : \; \sum_{i,j} \mathsf{d}^p(\xi^i, \xi^j) \boldsymbol{\gamma}_{ij} \le \theta^p, \; \sum_j \boldsymbol{\gamma}_{ij} = p_i, \forall i, \; \sum_i \boldsymbol{\gamma}_{ij} = q_j, \forall j \right\}. \tag{2.20}$$

Proof. Proof. Reformulation (2.19) follows from Theorem 2.1, and (2.20) can be obtained

using the equivalent definition of Wasserstein distance in Example 2.2. □

2.5 Applications

In this section, we apply our results to on/off system control, intensity estimation and worst-case Value-at-Risk analysis. In the first two problem, the nominal distribution is a point process, and the corresponding underlying space Ξ is the space of counting measures (sample paths), which is non-convex and infinite dimensional. In the third problem, the nonimal distribution is arbitrary probability distribution on a finite dimensional space, such as Gaussian distribution. Hence, the results in [24] and [25] cannot be applied.

2.5.1 On/Off System Control

In this problem, the decision maker faces a point process and controls a two-state (on/off) system. The point process is assumed to be exogenous, that is, the arrival times are not affected by the on/off state of the system. When the system is switched on, a cost of c per unit time is incurred, and each arrival while the system is on contributes 1 unit revenue. When the system is off, no cost is incurred and no revenue is earned. The decision maker wants to choose a control to maximize the total profit during a finite time horizon. This problem is a prototype for problems in sensor network and revenue management.

In many practical settings, the decision maker does not have a probability distribution for the point process. Instead, the decision maker has observations of historical sample paths of the point process, which constitute an empirical point process. Note that if one would use the Sample Average Approximation (SAA) method with the empirical point process, it would yield a degenerate control, in which the system is switched on only at the arrival time points of the empirical point process. Consequently, if future arrival times can differ from the empirical arrival times by even a little bit, the system would be switched off and no revenue would be earned. Due to such degeneracy and instability of the SAA method, we resort to the distributionally robust approach.

To model the problem, we scale the finite time horizon to $[0,1]$, and let

$$\Xi = \Big\{ \sum_{m=1}^{M} \boldsymbol{\delta}_{\xi_m} : M \in \mathbb{Z}_+,\, \xi_m \in [0,1],\, m = 1, \ldots, M \Big\}$$

be the space of finite counting measures on $[0,1]$. Then the point processes on $[0,1]$ are then defined by the set $\mathcal{P}(\Xi)$ of Borel probability measures on Ξ. To define the Wasserstein distance between to point processes $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\Xi)$, we need to define the metric d on the space Ξ of counting measures. We assume that the metric d on Ξ satisfies the following conditions (note that in this subsection, when we write the $\mathcal{W}_1$ distance between two Borel measures, we use the extended definition mentioned in Section 2.3):

(i) The metric space (Ξ, d) is a Polish space.

(ii) For any $\widehat{\boldsymbol{\eta}} = \sum_{m=1}^{M} \boldsymbol{\delta}_{\zeta_m}$ and $\boldsymbol{\eta} = \sum_{m=1}^{M} \boldsymbol{\delta}_{\xi_m}$, where m is a nonnegative integer and $\{\zeta_m\}_{m=1}^{M}, \{\xi_m\}_{m=1}^{M} \subset [0,1]$, it holds that

$$\mathsf{d}(\boldsymbol{\eta}, \widehat{\boldsymbol{\eta}}) = \mathcal{W}_1(\boldsymbol{\eta}, \widehat{\boldsymbol{\eta}}) = \sum_{m=1}^{M} |\xi_{(m)} - \zeta_{(m)}|,$$

where $\xi_{(m)}$ (resp. $\zeta_{(m)}$) are the order statistics of ξ_m (resp. ζ_m).

(iii) For any Borel set $C \subset [0,1]$, $\bar{\theta} \geq 0$, and $\widehat{\boldsymbol{\eta}} = \sum_{m=1}^{M} \boldsymbol{\delta}_{\zeta_m}$, where M is a positive integer and $\{\zeta_m\}_{m=1}^{M} \subset [0,1]$, it holds that

$$\inf_{\boldsymbol{\eta} \in \Xi} \big\{ \boldsymbol{\eta}(C) \,:\, \mathsf{d}(\boldsymbol{\eta}, \widehat{\boldsymbol{\eta}}) = \bar{\theta} \big\} \geq \inf_{\tilde{\boldsymbol{\eta}} \in \mathcal{B}([0,1])} \Big\{ \tilde{\boldsymbol{\eta}}(C) \,:\, \mathcal{W}_1(\tilde{\boldsymbol{\eta}}, \widehat{\boldsymbol{\eta}}) \leq \bar{\theta} \Big\}.$$

We note that condition (ii) is only imposed on $\boldsymbol{\eta}, \widehat{\boldsymbol{\eta}} \in \Xi$ such that $\boldsymbol{\eta}([0,1]) = \widehat{\boldsymbol{\eta}}([0,1])$. Possible choices for d are

$$\mathsf{d}\Big(\sum_{m=1}^{M} \boldsymbol{\delta}_{\xi_m}, \sum_{l=1}^{L} \boldsymbol{\delta}_{\zeta_l} \Big) = \sum_{m=1}^{\min\{M,L\}} \big| \xi_{(m)} - \zeta_{(l)} \big| + |M - L|\,,$$

$$\mathsf{d}\Big(\sum_{m=1}^{M}\boldsymbol{\delta}_{\xi_m},\sum_{l=1}^{L}\boldsymbol{\delta}_{\zeta_l}\Big)=\begin{cases}\max\{M,L\}, & M\neq L,\\ \sum_{m=1}^{M}\left|\xi_{(m)}-\zeta_{(m)}\right|, & M=L,\end{cases}$$

or

$$\mathsf{d}\Big(\sum_{m=1}^{M}\boldsymbol{\delta}_{\xi_m},\sum_{l=1}^{L}\boldsymbol{\delta}_{\zeta_l}\Big)=\begin{cases}+\infty, & M\neq L,\\ \sum_{m=1}^{M}\left|\xi_{(m)}-\zeta_{(m)}\right|, & M=L.\end{cases} \tag{2.21}$$

These metrics are similar to the ones in [78] and [79]. Given the metric d, we choose the distance between two point processes $\boldsymbol{\mu},\boldsymbol{\nu}\in\mathcal{P}(\Xi)$ to be $\mathcal{W}_1(\boldsymbol{\mu},\boldsymbol{\nu})$ as defined in (4.1).

Suppose we have n sample paths $\widehat{\boldsymbol{\eta}}^i=\sum_{m=1}^{M_i}\boldsymbol{\delta}_{\widehat{\xi}^i_m}$, $i=1,\ldots,n$, where M_i is a nonnegative integer and $\widehat{\xi}^i_m\in[0,1]$ for all i,m. Then the nominal distribution $\boldsymbol{\nu}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}_{\widehat{\boldsymbol{\eta}}^i}$, and the ambiguity set $\mathfrak{M}=\{\boldsymbol{\mu}\in\mathcal{P}(\Xi)\,:\,\mathcal{W}_1(\boldsymbol{\mu},\boldsymbol{\nu})\leq\theta\}$. Let $\mathcal{D}$ denote the set of all functions $\beta:[0,1]\to\{0,1\}$ such that $\beta^{-1}(1)$ is a Borel set, where $\beta^{-1}(1):=\{t\in[0,1]\,:\,\beta(t)=1\}$. The decision maker is looking for a control $\beta\in\mathcal{D}$ that maximizes the total profit, by solving the problem

$$v^*\;:=\;\sup_{\beta\in\mathcal{D}}\left\{v(\beta)\;:=\;-c\int_0^1\beta(t)dt+\inf_{\boldsymbol{\mu}\in\mathfrak{M}}\mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}\big[\boldsymbol{\eta}(\beta^{-1}(1))\big]\right\}. \tag{2.22}$$

We now investigate the structure of the optimal control. Let $\mathrm{int}(\beta^{-1}(1))$ be the interior of the set $\beta^{-1}(1)$ on the space $[0,1]$ with canonical topology (and thus $0,1\in\mathrm{int}([0,1])$).

Proposition 2.4. *For any $\boldsymbol{\nu}\in\mathcal{P}(\Xi)$ and control β, it holds that*

$$\begin{aligned}&\inf_{\boldsymbol{\mu}\in\mathfrak{M}}\mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}[\boldsymbol{\eta}(\beta^{-1}(1))]\\ =&\inf_{\rho\in\mathcal{P}(\mathcal{B}([0,1])\times\Xi)}\Big\{\mathbb{E}_{(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\sim\rho}[\boldsymbol{\eta}(\mathrm{int}(\beta^{-1}(1)))]:\;\mathbb{E}_{(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\sim\rho}\big[\mathcal{W}_1(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\big]\leq\theta,\;\pi^2_{\#}\rho=\boldsymbol{\nu}\Big\},\end{aligned} \tag{2.23}$$

Suppose $\boldsymbol{\nu}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}_{\widehat{\boldsymbol{\eta}}^i}$ with $\widehat{\boldsymbol{\eta}}^i=\sum_{m=1}^{M_i}\boldsymbol{\delta}_{\widehat{\xi}^i_m}$. There exists a non-negative integer M

such that

$$v^* = \sup_{\substack{\underline{\beta}_j,\overline{\beta}_j\in[0,1],\\ \underline{\beta}_j<\overline{\beta}_j<\underline{\beta}_{j'}<\overline{\beta}_{j'}\\ \forall 1\le j<j'\le M}} \Big\{ v\big(\sum_{j=1}^M \mathbb{1}_{[\underline{\beta}_j,\overline{\beta}_j]}\big) := -c\sum_{j=1}^M(\overline{\beta}_j-\underline{\beta}_j) + \inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}\big[\boldsymbol{\eta}\{\cup_{j=1}^M[\underline{\beta}_j,\overline{\beta}_j]\}\big] \Big\}. \tag{2.24}$$

Note that

$$\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\mu}}[\boldsymbol{\eta}(\beta^{-1}(1))] = \inf_{\boldsymbol{\gamma}\in\mathcal{P}(\Xi^2)} \big\{ \mathbb{E}_{(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\sim\boldsymbol{\gamma}}[\boldsymbol{\eta}(\beta^{-1}(1))] :\ \mathbb{E}_{\boldsymbol{\gamma}}[\mathsf{d}(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})]\le\theta,\ \pi^2_{\#}\boldsymbol{\gamma}=\boldsymbol{\nu}\big\}.$$

Hence (2.23) shows that without changing the optimal value, we can replace d by $\mathcal{W}_1$ in the constraint, and enlarge the set of joint distributions from $\mathcal{P}(\Xi^2)$ to $\mathcal{P}(\mathcal{B}([0,1])\times\Xi)$. Moreover, (2.24) shows that it suffices to consider the set of polices of which the duration of on-state is a finite disjoint union of intervals with positive length. We next show that given a control $\sum_{j=1}^M \mathbb{1}_{[\underline{\beta}_j,\overline{\beta}_j]}$, the computation of worst-case point process reduces to a linear program. For every $1\le i\le n$ and $1\le m\le M_i$, if $\widehat{\xi}^i_m \in \cup_{j=1}^M[\underline{\beta}_j,\overline{\beta}_j]$, we set $j^i_m\in\{1,\dots,M\}$ to be such that $\widehat{\xi}^i_m\in[\underline{\beta}_{j^i_m},\overline{\beta}_{j^i_m}]$, otherwise $j^i_m=0$. We also set x_0 to be any real number.

Proposition 2.5. *The objective $v\big(\sum_{j=1}^M \mathbb{1}_{[\underline{\beta}_j,\overline{\beta}_j]}\big)$ defined in (2.24) can be written as*

$$\sum_{j=1}^M -c(\overline{\beta}_j-\underline{\beta}_j) + \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{[\underline{\beta}_j,\overline{\beta}_j]}(\widehat{\xi}^i_m) + \min_{\substack{\underline{p}^i_m,\overline{p}^i_m\ge 0\\ \underline{p}^i_m+\overline{p}^i_m\le 1}} \Big\{ -\frac{1}{n}\sum_{i=1}^n \sum_{1\le m\le M_i: j^i_m>0} (\underline{p}^i_m+\overline{p}^i_m) :$$
$$\frac{1}{n}\sum_{i=1}^n \sum_{1\le m\le M_i: j^i_m>0} \big(\underline{p}^i_m|\underline{\beta}_{j^i_m}-\widehat{\xi}^i_m| + \overline{p}^i_m|\overline{\beta}_{j^i_m}-\widehat{\xi}^i_m|\big)\le\theta \Big\}.$$

Moreover, the above linear program can be solved by a greedy algorithm (see Algorithm

1), and there exists a worst-case point process that has the form

$$\boldsymbol{\mu}^*(x) = \frac{1}{n}\sum_{\substack{i=1\\ i\neq i_0}}^{n} \boldsymbol{\delta}_{\boldsymbol{\eta}_*^i} + \frac{p_0}{n}\boldsymbol{\delta}_{\underline{\boldsymbol{\eta}}_*^{i_0}} + \frac{(1-p_0)}{n}\boldsymbol{\delta}_{\overline{\boldsymbol{\eta}}_*^{i_0}},$$

where $i_0 \in \{1,\ldots,n\}$, $\boldsymbol{\eta}_*^i, \in \Xi$, $\boldsymbol{\eta}_*^i([0,1]) = \widehat{\boldsymbol{\eta}}^i([0,1])$ *for all* $i \neq i_0$, $\underline{\boldsymbol{\eta}}_*^{i_0}, \overline{\boldsymbol{\eta}}_*^{i_0} \in \Xi$ *and* $\underline{\boldsymbol{\eta}}_*^{i_0}([0,1]) = \overline{\boldsymbol{\eta}}_*^{i_0}([0,1]) = \widehat{\boldsymbol{\eta}}^{i_0}([0,1])$.

Algorithm 1 Greedy Algorithm

1: $\bar{\theta} \leftarrow 0$. $k \leftarrow 1$. $\overline{p}_m^i, \underline{p}_m^i \leftarrow 0$, $d_m^i \leftarrow \min(|\underline{\beta}_{\underline{j}_M^i} - \widehat{\xi}_m^i|, |\overline{\beta}_{\underline{j}_M^i} - \widehat{\xi}_m^i|), \forall i, m$.
2: Sort $\{d_m^i\}_{1\le i\le n, 1\le m\le M_i}$ in increasing order, denoted by $d_{m_{(1)}}^{i_{(1)}} \le d_{m_{(2)}}^{i_{(2)}} \le \ldots \le d_{m_{(\sum_{i=1}^n M_i)}}^{i_{(\sum_{i=1}^n m_i)}}$.
3: **while** $\bar{\theta} < n\theta$ **do**
4: **if** $d_t^i = |\underline{\beta}_{\underline{j}_m^i} - \widehat{\xi}_m^i|$ **then** $\underline{p}_{m_{(k)}}^{i_{(k)}*} \leftarrow \min\left(1, (n\theta - \bar{\theta})/d_{m(k)}^{i_{(k)}}\right)$.
5: **else** $\overline{p}_{m_{(k)}}^{i_{(k)}*} \leftarrow \min\left(1, (n\theta - \bar{\theta})/d_{m_{(k)}}^{i_{(k)}}\right)$.
6: **end if**
7: $k \leftarrow k+1$.
8: **end while**

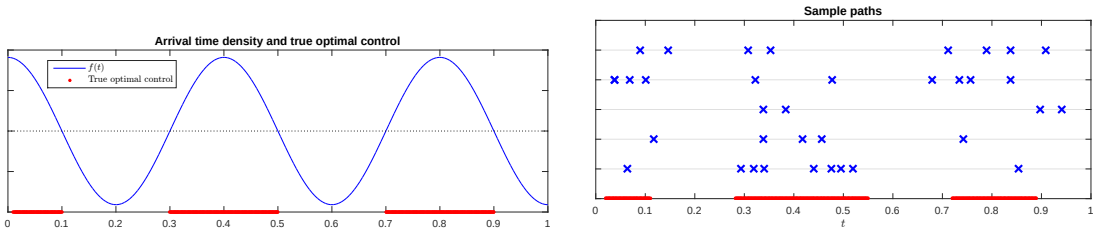


Figure 2.4: Optimal control for the true process and the DRSO

Example 2.8. We illustrate our results as follows. Suppose the number of arrivals has Poisson distribution $Poisson(\lambda)$, and given the number of arrivals, the arrival times are i.i.d. with density $f(t)$, $t \in [0,1]$. Then problem (2.22) is $\max_\beta \int_{\beta^{-1}(1)}[-c + \lambda f(t)]dt$, with optimal control $\beta^*(t) = 1_{\{\lambda f(t) > c\}}$. Note that $f \equiv 1$ corresponds to the Poisson point process with rate λ. In this example, we instead consider $f(t) = k[a + \sin(wt + s)]$, with $a > 1$ and $k = 1/[a + (\cos(s) - \cos(w+s))/w]$. Particularly, let $w = 5\pi$, $s = \frac{5}{2}\pi$, $a = 1.1$ and $c = \lambda = 10$. Thus ${\beta^*}^{-1}(1) = [0, 0.1] \cup [0.3, 0.5] \cup [0.7, 0.9]$. In the

numerical experiment, suppose we have $N = 10$ sample paths, each of which contains $M_i \sim Poisson(\lambda)$, $i = 1, \ldots, n$, i.i.d. arrival time points. The optimal controls for the true process and the DRSO are shown in Figure 2.4. We observe that even with a relatively small number of samples, the two controls differ from each other not too much, and thus the DRSO indeed provides a good solution to the original process control problem.

2.5.2 Intensity Estimation for Non-homogeneous Poisson Process

Consider estimating the intensity function $\beta(t)$ of a non-homogeneous Poisson process $A(t)$ using maximum likelihood method. Given n i.i.d. sample paths $\widehat{\boldsymbol{\eta}}^i = \sum_{m=1}^{M_i} \boldsymbol{\delta}_{\widehat{\xi}^i_m}$, $i = 1, \ldots, n$, the log-likelihood function (see, e.g. [80]) is written as

$$\int_0^T -\beta(t)dt + \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M_i} -\ln(\beta(\widehat{\xi}^i_m)).$$

A common practice is to partition the time horizon $[0, T]$ into several intervals, and assume $\beta(t)$ is piecewise constant on each interval. Then the maximum likelihood estimator equals to the average arrival rate on each interval. Such a approach suffers from the drawback that the estimator is sensitive to the partition of intervals. If the partition is so fine that many intervals may have zero observations, then the estimator also vanishes on these intervals. On the other hand, if the partition is very coarse, then the estimator remains constant during a long interval, which may not reflect the reality. It appears that there is no systematic way to adaptively choose the partition for this sample average method. Meanwhile, distributionally robust formulation with ϕ-divergence has the same problem, since the yielding estimator vanishes on intervals with zero observation.

Consider the distributionally robust formulation with Wasserstein distance

$$\min_{\beta} \left\{ \int_0^T \beta(t)dt + \max_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}\Big[\int_0^T -\ln(\beta(t))\boldsymbol{\eta}(dt)\Big] \right\}, \tag{2.25}$$

where $\mathfrak{M}$ is the same as the one in the previous subsection, namely, the Wasserstein ball

centered at the empirical process. To facilitate further analysis, we choose (2.21) as the definition of distance between two counting measures. Our strong duality results imply that the dual reformulation of (2.25) is given by

$$\min_{\substack{\beta \\ \lambda \geq 0}} \left\{ \int_0^T \beta(t)dt + \lambda\theta + \frac{1}{n}\sum_{i=1}^{n} \sum_{1 \leq m \leq M_i} \sup_{\xi \in [0,T]} \left\{ -\ln(\beta(\xi)) - \lambda|\xi - \widehat{\xi}_m^i| \right\} \right\}.$$

The following proposition suggests that the optimal estimator is constant if the radius of Wasserstein ball goes to infinity.

Proposition 2.6. *For sufficiently large θ, the optimal value $\beta_*(t)$ is constant.*

To numerically solve the problem, let us assume $\beta(t)$ is piecewise constant. In our numerical experiments, we assume the underlying true intensity function is given by $\beta(t) = 0.5 + 0.5t$ or $\beta(t) = 1 + \sin(\pi t)$, $t \in [0, 10]$. We fix the sample size (number of sample paths) $N = 20$ and vary the number of pieces in $\{20, 50, 100\}$. The radius θ is chosen via cross-validation method, for which half of the sample paths are used for training and the remaining are used for calibration. The out-of-sample performance is measured in terms of L_2 distance between the estimated intensity and true intensity. The estimation results and out-of-sample performances are shown in Figure 2.5 and Table 2.1. We observe that DRSO with Wasserstein distance has superior out-of-sample performance in all cases. The shape of the estimator from DRSO is insensitive to the fineness of the partition for the piecewise constant function. In contrast, the maximum likelihood estimator behaves terribly if we do not make the partition correctly, for example, when the number of pieces is too large.

Table 2.1: Out-of-sample performance of DRSO and SAA

	3c—$0.2 + 0.2t$			3c1 $+ \sin(\pi t)$		
Pieces	20	50	100	20	50	100
Wasserstein	0.394	0.481	0.544	2.008	2.122	2.276
SAA	1.510	6.536	11.906	6.160	6.591	11.766

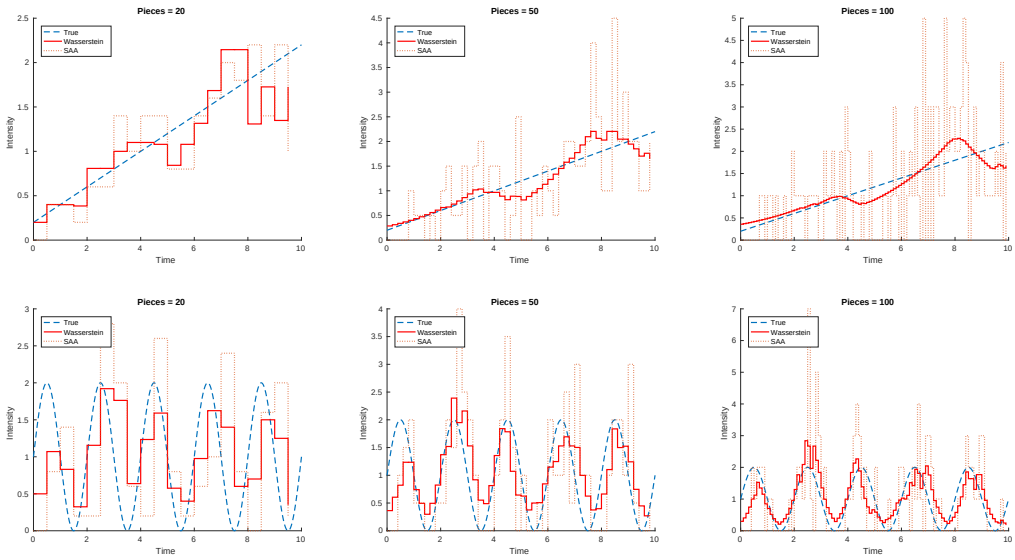


Figure 2.5: Estimation of intensity function using DRSO and SAA

2.5.3 Worst-case Value-at-Risk

Value-at-risk is a popular risk measure in financial industry. Given a random variable z and $\alpha \in (0,1)$, the value-at-risk $\mathrm{VaR}_\alpha[z]$ of z with respect to measure $\boldsymbol{\nu}$ is defined by

$$\mathrm{VaR}_\alpha[z] := \inf\left\{t : \mathbb{P}_{\boldsymbol{\nu}}\{z \le t\} \ge 1-\alpha\right\}.$$

In the spirit of [15], we consider the following worst-case VaR problem. Suppose we are given a portfolio consisting of n assets with allocation weight β satisfying $\sum_{k=1}^K \beta_k = 1$ and $\beta \ge 0$. Let $\boldsymbol{\xi}_k$ be the (random) return rate of asset k, $k = 1, \ldots, K$, and $r = \mathbb{E}[\boldsymbol{\xi}]$ the vector of the expected return rates. Assume the metric $\mathsf{d}(\cdot,\cdot)$ is induced by the infinity norm $||\cdot - \cdot||_\infty$ on $\mathbb{R}^K$. The worst-case VaR with respect to the set of probability distributions $\mathfrak{M}$ is defined as

$$\mathrm{VaR}_\alpha^{wc}(\beta) := \min_q \left\{ q : \inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{P}_{\boldsymbol{\mu}}\{-\beta^\top \boldsymbol{\xi} \le q\} \ge 1-\alpha \right\}.$$

Proposition 2.7. *Let* $q \ge \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]$, $\theta > 0$, $\alpha \in (0,1)$, $\beta \in \{\beta' \in \mathbb{R}^K : \sum_{k=1}^K \beta'_k =$

$1,\ \beta' \geq 0\}$. *Define*

$$\alpha_0 := \min\left(1, \frac{(\alpha - \boldsymbol{\nu}\{\xi : -\beta^\top \xi > \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\})(q - \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}])^p}{\left|\theta^p - \mathbb{E}_{\boldsymbol{\nu}}\left[(q+\beta^\top \xi)^p \mathbb{1}_{\{-\beta^\top \xi > \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\}}\right]\right|}\right).$$

Then $\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{P}_{\boldsymbol{\mu}}\{-\beta^\top \boldsymbol{\xi} \leq q\} \geq 1-\alpha$ *is equivalent to*

$$\mathbb{E}_{\boldsymbol{\nu}}\left[((q+\beta^\top \boldsymbol{\xi})^+)^p + \cdot \mathbb{1}_{\{-\beta^\top \boldsymbol{\xi} > \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\}}\right] + \alpha_0 \mathbb{E}_{\boldsymbol{\nu}}\left[((q+\beta^\top \boldsymbol{\xi})^+)^p \cdot \mathbb{1}_{\{-\beta^\top \boldsymbol{\xi} = \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\}}\right] \geq \theta^p.$$

In particular, when $\boldsymbol{\nu}$ *is a continuous distribution, the condition above can be reduced to*

$$\mathbb{E}_{\boldsymbol{\nu}}\left[((q+\beta^\top \boldsymbol{\xi})^+)^p \mathbb{1}_{\{-\beta^\top \boldsymbol{\xi} \geq \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\}}\right] \geq \theta^p.$$

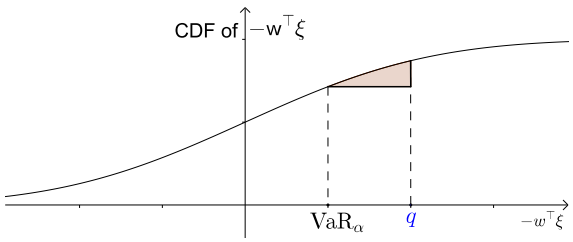


Figure 2.6: Worst-case VaR. When $-\beta^\top \boldsymbol{\xi}$ is continuously distributed and $p = 1$, VaR_α^{wc} equals to the q such that the area of the shade region is equal to θ.

Example 2.9 (Worst-case VaR **with Gaussian nominal distribution).** Suppose $\boldsymbol{\nu} \sim Normal(\mu, \Sigma)$ and consider $p = 1$. It follows that $-\beta^\top \boldsymbol{\xi} \sim Normal(-\beta^\top \mu, \beta^\top \Sigma \beta)$ and $\mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}] = -\beta^\top \mu + \sqrt{\beta^\top \Sigma \beta} \cdot \Phi^{-1}(1-\alpha)$. By Proposition 2.7, $\mathrm{VaR}_\alpha^{wc}(\beta)$ is the minimal q such that (see Figure 2.6)

$$f(q) := \frac{1}{\sqrt{2\pi \beta^\top \Sigma \beta}} \int_{\mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]}^{q} (q-y) e^{-\frac{(y+\beta^\top \mu)^2}{2\beta^\top \Sigma \beta}} dy \geq \theta. \tag{2.26}$$

Since $f(q)$ is monotone, (2.26) can be solved efficiently via any one-dimensional search algorithm.

We remark that the above result indicates that finding the worst-case VaR is tractable.

It should be noted that, however, finding the best allocation weight, i.e., optimizing over w is still hard, since the VaR constraint is essentially a chance-constraint.

2.6 Discussions

In this section, we discuss some advantages of Wasserstein ambiguity set. In Section 2.6.1, we compare the Wasserstein ambiguity set to ϕ-divergence ambiguity set for newsvendor problem. In Section 2.6.2, we illustrate how the close connection between Wasserstein DRSO and robust programming (Corollary 2.3(iii)) can expand the tractability of Wasserstein DRSO.

2.6.1 Newsvendor problem: a comparison to ϕ-divergence

In this subsection, we discuss some advantages of Wasserstein ambiguity set by performing a numerical study on distributionally robust newsvendor problems, with an emphasis on the worst-case distribution.

In the newsvendor model, the decision maker has to decide the inventory level before the random demand is realized, facing both overage and underage costs. The problem can be formulated as

$$\min_{\beta \geq 0} \mathbb{E}_{\boldsymbol{\mu}}[h(\beta - \boldsymbol{\xi})^+ + b(\boldsymbol{\xi} - \beta)^+],$$

where β is the decision variable for initial inventory level, $\boldsymbol{\xi}$ is the random demand, and h, b represent respectively the overage and underage costs per unit. We assume that $b, h > 0$, and $\boldsymbol{\xi}$ is supported on $\{0, 1, \ldots, \bar{B}\}$ for some positive integer $\bar{B}$. Sometimes the demand data is expensive to obtain. For instance, a company is introducing a new product of which the demand data is collected by setting up pilot stores. Then the decision maker may want to consider the DRSO counterpart

$$\min_{\beta \geq 0} \sup_{\boldsymbol{\mu} \in \Delta_{\bar{B}}} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[h(\beta - \xi)^+ + b(\xi - \beta)^+] : \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \theta \right\}.$$

Table 2.2: Examples of ϕ-divergence

Kullback-Leibler	Burg entropy	χ^2-distance	Modified χ^2	Hellinger	Total Variation
ϕ_{kl}	ϕ_b	ϕ_{χ^2}	$\phi_{m\chi^2}$	ϕ_h	ϕ_{tv}
$\phi(t), t \geq 0$	$t \log t$	$-\log t$	$\frac{1}{t}(t-1)^2$	$(t-1)^2$	$(\sqrt{t}-1)^2$
$\lvert t-1 \rvert$					
$\sum p_j \log\left(\frac{p_j}{q_j}\right)$	$\sum q_j \log\left(\frac{q_j}{p_j}\right)$	$\sum \frac{(p_j-q_j)^2}{p_j}$	$\sum \frac{(p_j-q_j)^2}{q_j}$	$\sum(\sqrt{p_j}-\sqrt{q_j})^2$	$\sum \lvert p_j - q_j \rvert$

Using Corollary 2.3, we obtain a convex programming reformulation

$$\min_{\beta,\lambda\geq 0}\left\{\lambda\theta^p + \sum_{i=0}^{\bar{B}} q_i y_i :\ y_i \geq \max\left[h(\beta-j), b(j-\beta)\right] - \lambda|i-j|^p,\ \forall 0 \leq i,j \leq \bar{B}\right\}.$$

On the other hand, one may would also consider ϕ-divergence ambiguity set (Table 2.2 shows some common ϕ-divergences). As mentioned in Section 1.2.1, the worst-case distribution in ϕ-divergence ambiguity set may be problematic. Indeed, when $\lim_{t\to\infty}\phi(t)/t = \infty$, such as ϕ_{kl}, $\phi_{m\chi^2}$, the ϕ-divergence ambiguity set fails to include sufficiently many relevant distributions. In fact, since $0\phi(p_j/0) = p_j \lim_{t\to\infty}\phi(t)/t = \infty$ for all $p_j > 0$, the ϕ-divergence ambiguity set does not include any distribution which is not absolutely continuous with respect to the nominal distribution $\boldsymbol{\nu}$.

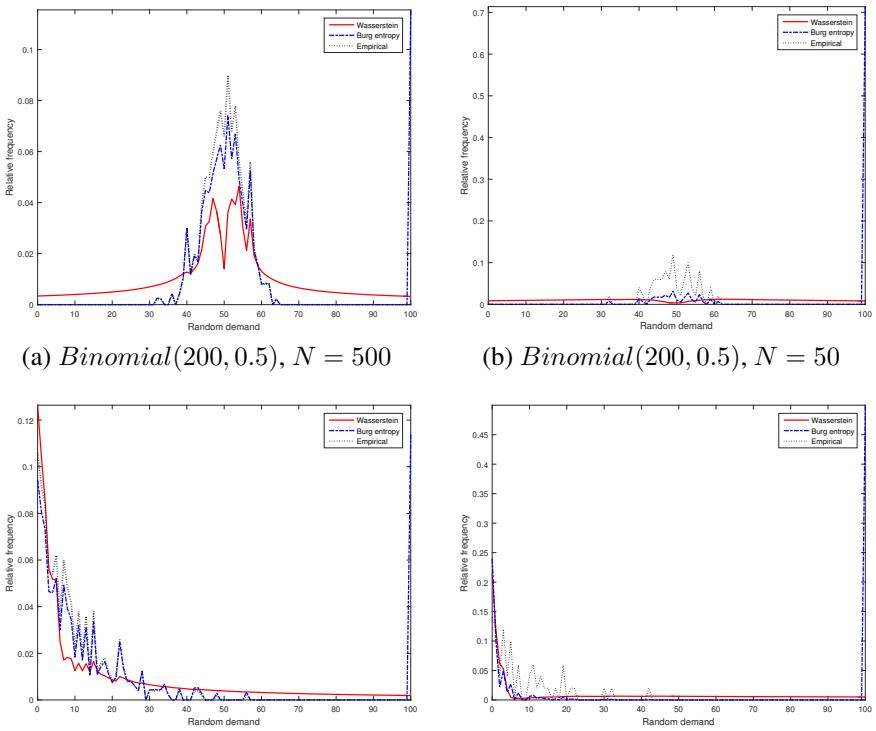
When $\lim_{t\to\infty}\phi(t)/t < \infty$, such as ϕ_b, ϕ_{χ^2}, ϕ_h, ϕ_{tv}, the situation is even worse. Define $I_0 := \{1 \leq j \leq n : q_j > 0\}$ and $j_M := \arg\max_{1\leq j\leq n}\{\ell(\xi^j) : q_j = 0\}$. Assume $\ell(\xi^j)$ are different from each other, then according to [18] and [17], the worst-case distribution satisfies

$$p_j^*/q_j \ \in\ \partial\phi^*\Big(\frac{\ell(\xi^j)-c^*}{\lambda^*}\Big),\ \forall i \in I_0, \tag{2.27a}$$

$$p_j^* \ =\ 0,\ \forall j \notin I_0 \cup \{j_M\}, \tag{2.27b}$$

$$p_{j_M}^* \ =\ \begin{cases} 1-\sum_{i\in I_0} p_j^*, & \text{if } c^* = \ell(\xi^{j_M}) - \lambda^* \lim_{t\to\infty}\phi(t)/t, \\ 0, & \text{if } c^* > \ell(\xi^{j_M}) - \lambda^* \lim_{t\to\infty}\phi(t)/t, \end{cases} \tag{2.27c}$$

for some $\lambda^* \geq 0$ and $c^* \geq \ell(\xi^{j_M}) - \lambda^* \lim_{t\to\infty}\phi(t)/t$. (2.27b) suggests that the support of the worst-case distribution and that of the nominal distribution can differ by at most



(a) $Binomial(200, 0.5)$, $N = 500$

(b) $Binomial(200, 0.5)$, $N = 50$

(c) truncated $Geometric(0.1)$, $N = 500$

(d) truncated $Geometric(0.1)$, $N = 50$

Figure 2.7: Histograms of worst-case distributions yielding from Wasserstein distance and Burg entropy

one point ξ^{j_M}. If $p^*_{j_M} > 0$, (2.27c) suggests that the probability mass is moved away from scenarios in I_0 to the worst scenario ξ^{j_M}. Note that in many applications where the support of $\boldsymbol{\xi}$ is unknown, the choice of the underlying space Ξ (and thus ξ^{j_M}) may be arbitrary. Hence the worst-case behavior is sensitive to the specification of Ξ and the shape of function ℓ.

We perform a numerical test of which setup is similar to [11] and [18]. We set $b = h = 1$, $\bar{B} = 100$, and $N \in \{50, 500\}$ representing small and large datasets. The data is then generated from $Binomial(100, 0.5)$ and $Geometric(0.1)$ truncated on $[0, 100]$. For a fair comparison, we estimate the radius of the ambiguity set such that it covers the underlying distribution with probability greater than 95%.

When the underlying distribution is Binomial, intuitively, the symmetry of Binomial distribution and $b = h = 1$ implies that the optimal initial inventory level is close to

$\bar{B}/2 = 50$, and the corresponding worst-case distribution should be similar to a mixture distribution with two modes, representing high and low demand respectively. This intuition is consistent with the solid curves in Figure (2.7a)(2.7b), representing the worst-case distribution in Wasserstein ambiguity set. In addition, their tail distributions are smooth and reasonable for both small and large datasets. In contrast, if Burg entropy is used to define the ambiguity set (dashed curves in Figure (2.7a)(2.7b)), the worst-case distribution has disconnected support, and is not symmetric. There is a large spike on the boundary 100, corresponding to the "popping" behavior mentioned in [17]. Especially when the dataset is small, the spike is huge, which makes the solution too conservative.

When the underlying distribution is Geometric, intuitively, the worst-case distribution should have one spike for low demand and a heavy tail for high demand. Again, this is consistent with the worst-case distribution in Wasserstein ambiguity set (solid curves in Figure (2.7c)(2.7d)). While using Burg entropy (dashed curves in Figure (2.7c)(2.7d)), the tail has unrealistic spikes on the boundary. For distribution with unbounded support, the tail distribution is very sensitive to our choice of truncation threshold $\bar{B}$. Hence, the conclusion for this numerical test is that Wasserstein ambiguity set is likely to yield a more reasonable, robust and realistic worst-case distribution.

2.6.2 Two-stage DRSO: connection with robust optimization

In Corollary 2.2(iii) we established the close connection between the DRSO problem and robust programming. More specifically, we show that every DRSO problem can be approximated by robust programs with rather high accuracy, which significantly enlarges the applicability of the DRSO problem. To illustrate this point, in this section we show the tractability of the two-stage linear DRSOs.

Consider the two-stage distributionally robust stochastic optimization

$$\min_{\beta \in \mathcal{D}} c^\top \beta + \sup_{\boldsymbol{\mu} \in \mathfrak{M}} \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta, \boldsymbol{\xi})], \tag{2.28}$$

where $\ell(\beta, \boldsymbol{\xi})$ is the optimal value of the second-stage problem

$$\min_{\chi \in \mathbb{R}^m} \left\{ q(\boldsymbol{\xi})^\top \chi : T(\boldsymbol{\xi})\beta + W(\boldsymbol{\xi})\chi \leq h(\boldsymbol{\xi}) \right\},$$

and

$$q(\boldsymbol{\xi}) = q^0 + \sum_{l=1}^{s} \boldsymbol{\xi}_l q^l,\ T(\boldsymbol{\xi}) = T^0 + \sum_{l=1}^{s} \boldsymbol{\xi}_l T^l,\ W(\boldsymbol{\xi}) = W^0 + \sum_{l=1}^{s} \boldsymbol{\xi}_l W^l,\ h(\boldsymbol{\xi}) = h^0 + \sum_{l=1}^{s} \boldsymbol{\xi}_l h^l.$$

We assume $p = 2$ and $\Xi = \mathbb{R}^K$ with Euclidean distance d. In general, the two-stage problem (2.28) is NP-hard. However, we are going to show that with tools from robust programming, we are able to obtain a tractable approximation of (2.28). Let $\mathfrak{M}_1 := \{(\xi^1, \ldots, \xi^n) \in \Xi^n : \frac{1}{n}\sum_{i=1}^n ||\xi^i - \widehat{\xi}^i||_2^2 \leq \theta^2\}$. Using Theorem 2.2(ii) with $K = 1$, we obtain an adjustable-robust-programming approximation

$$\begin{aligned}
&\min_{\beta \in \mathcal{D}} \left\{ c^\top \beta + \sup_{(\xi^i)_i \in \mathfrak{M}_1} \frac{1}{n} \sum_{i=1}^n \ell(\beta, \xi^i) \right\} \\
&= \min_{\substack{\beta \in \mathcal{D}, t \in \mathbb{R} \\ \chi : \Xi \to \mathbb{R}}} \left\{ t : \begin{array}{l} t \geq c^\top \beta + \frac{1}{n}\sum_i q(\xi^i)^\top \chi(\xi^i),\ \forall (\xi^i)_i \in \mathfrak{M}_1, \\ T(\xi)\beta + W(\xi)\chi(\xi) \leq h(\xi),\ \forall \xi \in \bigcup_{i=1}^n \{\xi' \in \Xi : ||\xi' - \widehat{\xi}^i||_2 \leq \theta\sqrt{n}\} \end{array} \right\},
\end{aligned} \tag{2.29}$$

where the second set of inequalities follows from the fact that $T(\xi)x + W(\xi)\chi(\xi) \leq h(\xi)$ should hold for any realization ξ with positive probability for some distribution in $\mathfrak{M}_1$. Although problem (2.29) is still intractable in general, there has been a substantial literature on different approximations to problem (2.29). One popular approach is to consider the so-called affinely adjustable robust counterpart (AARC) as follows. We assume that χ is an affine function of ξ:

$$\chi(\xi) = \chi^0 + \sum_{l=1}^{s} \xi_l \chi^l,\ \forall \xi \in \bigcup_{i=1}^{n} B^i,$$

for some $\chi^0, \chi^l \in \mathbb{R}^m$, where $B^i := \{\xi' \in \Xi : ||\xi' - \widehat{\xi}^i||_2 \leq \theta\sqrt{n}\}$. Then the AARC of

(2.29) is

$$\min_{\substack{\beta\in\mathcal{D},t\in\mathbb{R}\\ \chi^l\in\mathbb{R}^m,l=0,\ldots,s}}\Bigg\{t:\ c^\top\beta+\frac{1}{n}\sum_{i=1}^n\Big(q^0+\sum_{l=1}^s\xi_l^iq^l\Big)^\top\Big(\chi^0+\sum_{l=1}^s\xi_l^i\chi^l\Big)-t\le 0,\ \forall(\xi^i)_i\in\mathfrak{M}_1,$$
$$\Big(T^0+\sum_{l=1}^s\xi_lT^l\Big)\beta+\Big(W^0+\sum_{l=1}^s\xi_lW^l\Big)\Big(\chi^0+\sum_{l=1}^s\xi_l\chi^l\Big)-\Big(h^0+\sum_{l=1}^sh^l\xi_l\Big)\le 0,$$
$$\forall\xi\in\bigcup_{i=1}^nB^i\Bigg\}. \tag{2.30}$$

Set $\zeta_{il}:=\xi_l^i-\widehat{\xi}_l^i$ for $i=1,\ldots,n$ and $l=1,\ldots,s$. In view of $\mathfrak{M}_1$, ζ belongs to the ellipsoidal uncertainty set

$$\mathcal{U}_\zeta=\{(\zeta_{il})_{i,l}:\sum_{i,l}\zeta_{il}^2\le n\theta^2\}.$$

Set $\rho=[\beta;t;\{\chi^l\}_{l=0}^s]$, and define

$$\begin{cases}\alpha_0(\rho):=-\Big[c^\top\beta+\dfrac{1}{n}\sum\limits_{i=1}^n(q^0+\sum\limits_{l=1}^s\widehat{\xi}_l^iq^l)^\top(\chi^0+\sum\limits_{l=1}^s\widehat{\xi}_l^i\chi^l)-t\Big],\\ \alpha_0^{il}(\rho):=-\dfrac{\Big[(q^0+\sum_{l'=1}^s\widehat{\xi}_{l'}^iq^{l'})^\top\chi^{l'}+{q^{l'}}^\top(\chi^0+\sum_{l'=1}^s\widehat{\xi}_{l'}^i\chi^{l'})\Big]}{2N},\quad\forall 1\le i\le n,\ 1\le l,l'\le s.\\ \Gamma_0^{(l,l')}(\rho):=-\dfrac{q^l\chi^{l'}+q^{l'}\chi^l}{2n},\end{cases}$$

Then the first set of constraints in (2.30) is equivalent to

$$\alpha_0(\rho)+2\sum_{i,l}\alpha_0^{il}(\rho)\zeta_{il}+\sum_i\sum_{l,l'}\Gamma_0^{(l,l')}\zeta_{il}\zeta_{il'}\ge 0,\quad\forall(\zeta_{il})_{i,l}\in\mathcal{U}_\zeta. \tag{2.31}$$

It follows from Theorem 4.2 in [81] that (2.31) takes place if and only if there exists $\lambda_0\ge 0$ such that

$$(\alpha_0(\rho)-\lambda_0)v^2+2v\sum_{i,l}\alpha_0^{il}(\rho)w_{il}+\sum_i\sum_{l,l'}\Gamma_0^{(l,l')}w_{il}w_{i'l'}+\frac{\lambda_0}{n\theta^2}\sum_{i,l}w_{il}^2\ge 0,\ \forall v\in\mathbb{R},\forall w_{il}\in\mathbb{R},\forall i,l.$$

Or in matrix form,

$$\exists \lambda_0 \geq 0 : \begin{pmatrix} \Gamma_0 \otimes I_N + \frac{\lambda_0}{n\theta^2} \cdot I_{sN} & \mathrm{vec}(\alpha_0) \\ \mathrm{vec}^\top(\alpha_0) & \alpha_0(\rho) - \lambda_0 \end{pmatrix} \succeq 0, \tag{2.32}$$

where I_N (resp. I_{sN}) is n (resp. sN) dimensional identity matrix, $\otimes$ is the Kronecker product of matrices and vec is the vectorization of a matrix.

Now we reformulate the second set of constraints in (2.30). For all $1 \leq i \leq n$, $1 \leq j \leq m$ and $1 \leq l, l' \leq s$, we set

$$\begin{cases} \alpha_{ij}(\rho) := -\Big[(T_j^0 + \sum_{l=1}^s \widehat{\xi}_l^i T_j^l)\beta + (W_j^0 + \sum_{l=1}^s \widehat{\xi}_l^i W_j^l)(\chi^0 + \sum_{l=1}^s \widehat{\xi}_l^i \chi^l) - (h_j^0 + \sum_{l=1}^s \widehat{\xi}_l^i h_j^l)\Big], \\ \beta_{ij}^l(\rho) := -\dfrac{\Big[T_j^l\beta + (W_j^0 + \sum_{l=1}^s \widehat{\xi}_l^i W_j^l)\chi^l + W_j^l(\chi^0 + \sum_{l=1}^s \widehat{\xi}_l^i \chi^l) - h_l\Big]}{2}, \\ \Gamma_j^{(l,l')}(\rho) := -\dfrac{W_j^l \chi^{l'} + W_j^{l'} \chi^l}{2}. \end{cases}$$

Let $\eta^i := \xi - \widehat{\xi}^i$ for $1 \leq i \leq n$. Then the second set of constraints in (2.30) is equivalent to

$$\alpha_{ij}(\rho) + 2\beta_{ij}(\rho)^\top \eta^i + {\eta^i}^\top \Gamma_j(\rho) \eta^i \geq 0, \ \forall \eta^i \in \{\eta' \in \mathbb{R}^K : ||\eta'||_2 \leq \theta\sqrt{n}\}, \forall 1 \leq i \leq n, 1 \leq j \leq m.$$

Again by Theorem 4.2 in [81] we have further equivalence

$$\exists \lambda_{ij} \geq 0 : \begin{pmatrix} \Gamma_j(\rho) + \frac{\lambda_{ij}}{n\theta^2} \cdot I_s & \beta_{ij}(\rho) \\ \beta_{ij}(\rho)^\top & \alpha_{ij}(\rho) - \lambda_{ij} \end{pmatrix} \succeq 0, \quad \forall 1 \leq i \leq n, 1 \leq j \leq m. \tag{2.33}$$

Combining (2.32) and (2.33) we obtain the following result.

Proposition 2.8. *An exact reformulation of the AARC of (2.29) is given by*

$$\min_{\substack{\beta \in \mathcal{D}, t \in \mathbb{R}, \chi^l \in \mathbb{R}^m, l=1,\ldots s \\ \lambda_0, \lambda_{ij} \geq 0, i=1,\ldots,n, j=1,\ldots,s}} \{t : \ (2.32), (2.33) \textit{ holds } \}.$$

Note that (2.29) is a fairly good approximation of the original two-stage DRSO problem (2.28) by Theorem 2.1. Hence, as long as the AARC of (2.29) is reasonably good, its semidefinite-program reformulation (2.8) provides a good tractable approximation of the two-stage linear DRSO (2.28).

2.6.3 Distributionally robust transportation problem: an illustration of the constructive proof approach

In this subsection, we demonstrate the power of our proof method by applying the same idea to a class of distributionally robust transportation problems.

Suppose $\Xi \subset \mathbb{R}^2$ is bounded, and let A denote a Borel probability measure on Ξ. In the famous paper [82], it is shown that the length of the shortest traveling salesman tour through n i.i.d. random points with density f is asymptotically equal to $c\sqrt{n}\int_\Xi \sqrt{f}dA$ for some constant c. Since then, continuous approximations have been explored for many hard combinatorial problems, such as Steiner tree problems [83], space-filling curves [84, 85], facility location [86], and Steele's generalization to sub-additive Euclidean functionals [87, 88], which identifies a class of random processes whose limits are equal to $c\int_\Xi f^{(K-1)/K}dA$ for some c, where K is the dimension of Ξ.

Motivated by these results, [89] considers a continuous approximation of the traveling salesman problem in a distributionally robust setting. More specifically, they solve the worst-case problem $\sup_{f\in\mathfrak{A}}\int_\Xi \sqrt{f}dA$, in which the distributions with density functions f have to lie in a Wasserstein ball. Using duality theory for convex functionals, they are able to reformulate the problem and obtain a representation of the worst-case distribution.

In the same spirit, we consider a slightly more general problem as follows. Let

$$\begin{aligned}
\mathfrak{B} &:= \{d\boldsymbol{\mu}/dA \,:\, \boldsymbol{\mu} \in \mathcal{B}(\Xi),\, \boldsymbol{\mu} \text{ is absolutely continuous w.r.t } A\},\\
\mathfrak{P} &:= \{d\boldsymbol{\mu}/dA \,:\, \boldsymbol{\mu} \in \mathcal{P}(\Xi),\, \boldsymbol{\mu} \text{ is absolutely continuous w.r.t } A\},\\
\mathfrak{A} &:= \{d\boldsymbol{\mu}/dA \in \mathfrak{P} \,:\, \mathcal{W}_p(\boldsymbol{\mu},\boldsymbol{\nu}) \le \theta\},
\end{aligned}$$

where $d\boldsymbol{\mu}/dA$ denotes the Radon-Nikodym derivative. We use the overloaded notation $\mathcal{W}_p(f,\boldsymbol{\nu})$ to represent the distance $\mathcal{W}_p(\boldsymbol{\mu},\boldsymbol{\nu})$ between the nominal distribution $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^n \widehat{\xi}^i$ and the distribution $\boldsymbol{\mu} \in \mathcal{P}(\Xi)$ such that $f = d\boldsymbol{\mu}/dA$. Let $\mathcal{L} : \mathbb{R} \mapsto \mathbb{R}$ be an increasing concave function approaching infinity. Consider the problem

$$v_P = \sup_{f\in\mathfrak{A}} \int_\Xi \mathcal{L}\circ f\, dA. \tag{2.34}$$

Our goal is to derive the strong dual of (2.34) and obtain a representation for the worst-case distribution using the same proof method as in Section 2.4.1.

Step 1. Derive weak duality.

First, we derive weak duality by writing the Lagrangian and applying a similar reasoning to the proof of Proposition 2.1. Note that in Kantorovich's duality (2.2), the supremum can be restricted to $u, v \in C_b(\Xi)$ (cf. Section 1.3 of [70]). Then

$$\begin{aligned}
v_P &= \sup_{f\in\mathfrak{P}} \inf_{\lambda\geq 0} \left\{ \int_\Xi \mathcal{L}\circ f\, dA + \lambda(\theta^p - \mathcal{W}_p^p(f,\boldsymbol{\nu})) \right\} \\
&\leq \sup_{f\in\mathfrak{B}} \inf_{\lambda\geq 0} \left\{ \int_\Xi \mathcal{L}\circ f\, dA + \lambda(\theta^p - \mathcal{W}_p^p(f,\boldsymbol{\nu})) \right\} \\
&\leq \inf_{\lambda\geq 0} \left\{ \lambda\theta^p + \sup_{f\in\mathfrak{B}} \left\{ \int_\Xi \mathcal{L}\circ f\, dA - \lambda \mathcal{W}_p^p(f,\boldsymbol{\nu}) \right\} \right\} \\
&= \inf_{\lambda\geq 0} \left\{ \lambda\theta^p + \sup_{f\in\mathfrak{B}} \left\{ \int_\Xi \mathcal{L}\circ f\, dA \right.\right. \\
&\qquad \left.\left. -\lambda \sup_{u,v\in C_b(\Xi)} \left\{ \int_\Xi uf dA + \int_\Xi v d\boldsymbol{\nu} \;:\; u(\xi) \leq \inf_{\zeta\in\Xi} \mathsf{d}^p(\xi,\zeta) - v(\zeta) \right\} \right\} \right\} \\
&= \inf_{\lambda\geq 0} \left\{ \lambda\theta^p + \sup_{f\in\mathfrak{B}} \left\{ \int_\Xi \mathcal{L}\circ f\, dA \right.\right. \\
&\qquad \left.\left. -\lambda \sup_{v\in C_b(\Xi):\int_\Xi v d\boldsymbol{\nu}=0} \left\{ \int_\Xi \left[\inf_{\zeta\in\Xi} \mathsf{d}^p(\xi,\zeta) - v(\zeta) \right] f(\xi) A(d\xi) \right\} \right\} \right\} \\
&\leq \inf_{\substack{\lambda\geq 0 \\ v\in C_b(\Xi):\int_\Xi v d\boldsymbol{\nu}=0}} \left\{ \lambda\theta^p + \sup_{f\in\mathfrak{B}} \left\{ \int_\Xi [\mathcal{L}\circ f(\xi) - \lambda\Phi_v(\xi) f(\xi)] A(d\xi) \right\} \right\},
\end{aligned}$$

where the second inequality follows from Lemma 2.1, and in the last inequality $\Phi_v(\xi) :=$

$\inf_{\zeta\in\Xi}[\mathsf{d}^p(\xi,\zeta)-v(\zeta)]$. Let

$$\mathcal{L}^*(a) \;:=\; \sup_{t\geq 0}\mathcal{L}(t)-at, \quad a\in\mathbb{R},$$

which can be viewed as the Legendre transform of concave function $\mathcal{L}$. Thus $\mathcal{L}^*$ is convex and we denote by $\partial\mathcal{L}^*(a)$ its subdifferential at $a\in \text{dom}\mathcal{L}^*$, where $\text{dom}\mathcal{L}^* := \{a\geq 0 : \mathcal{L}^*(a)<\infty\}$. It follows that

$$\begin{aligned}
v_P \;&\leq \inf_{\substack{\lambda\geq 0\\ v\in C_b(\Xi):\int_\Xi v d\boldsymbol{\nu}=0}} \left\{\lambda\theta^p+\sup_{f\in\mathfrak{B}}\left\{\int_\Xi[\mathcal{L}\circ f(\xi)-\lambda\Phi_v(\xi)f(\xi)]A(d\xi)\right\}: \right.\\
&\qquad\qquad \left. \lambda\Phi_v(\xi)\in\text{dom}\mathcal{L}^*,\ A-a.e.\right\}\\
&\leq \inf_{\substack{\lambda\geq 0\\ v\in C_b(\Xi):\int_\Xi v d\boldsymbol{\nu}=0}} \left\{\lambda\theta^p+\int_\Xi \mathcal{L}^*(\lambda\Phi_v(\xi))A(d\xi)\right\}\\
&=: \inf_{\substack{\lambda\geq 0\\ v\in C_b(\Xi):\int_\Xi v d\boldsymbol{\nu}=0}} h_v(\lambda)\\
&=: v_D.
\end{aligned}$$

Step 2. Show the existence of a dual minimizer.

Since $\lim_{x\to\infty}\mathcal{L}(x)=\infty$, we have $(-\infty,0]\cap\text{dom}\mathcal{L}^*=\varnothing$. It follows that $\lambda\Phi_v>0$ and thus $\lambda>0$ and $v<\text{diam}(\Xi)$. Note that $\int_\Xi v d\boldsymbol{\nu}=0$, hence there exists $M>0$, such that $||v||_\infty<M$ for all feasible v, thereby Φ_v is bounded on Ξ uniformly in v. It follows that $h(\lambda)$ approaches to ∞ as $\lambda\to\infty$ uniformly in v. Using the fact that $\boldsymbol{\nu}=\frac{1}{n}\sum_{i=1}^n\widehat{\xi}^i$, we conclude that there exists $M>0$ such that

$$v_D=\inf_{\lambda,v}\left\{h(\lambda): 0\leq\lambda\leq M, |v(\widehat{\xi}^i)|\leq M, \sum_i v(\widehat{\xi}^i)=0\right\}.$$

Hence there exists a dual minimizer (λ^*,v^*).

Step 3. Use first-order optimality to construct a primal solution.

From Step 2 we know that $\lambda^* > 0$. The first-order optimality at λ^* reads

$$\begin{aligned} \theta^p + \frac{\partial}{\partial \lambda -} \int_\Xi \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) \Phi_{v^*}(\xi) A(d\xi) \leq 0, \\ \theta^p + \frac{\partial}{\partial \lambda +} \int_\Xi \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) \Phi_{v^*}(\xi) A(d\xi) \geq 0. \end{aligned} \tag{2.35}$$

Since Ξ is bounded, it follows that $\partial \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi))$ is bounded on Ξ, thus we can exchange differentiation and integration operators in the inequalities above. We define

$$f^*(\xi) := -\big[p^* \frac{\partial}{\partial \lambda -} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) + (1 - p^*) \frac{\partial}{\partial \lambda +} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi))\big], \ \forall \xi \in \operatorname{supp} A, \tag{2.36}$$

where

$$p^* := \frac{\theta^p + \int_\Xi \frac{\partial}{\partial \lambda +} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) \Phi_{v^*}(\xi) A(d\xi)}{\int_\Xi \frac{\partial}{\partial \lambda +} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) \Phi_{v^*}(\xi) A(d\xi) - \int_\Xi \frac{\partial}{\partial \lambda -} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)) \Phi_{v^*}(\xi) A(d\xi)},$$

provided that the denominator is nonzero, otherwise we set $p^* = 1$. By definition of $\mathcal{L}^*$, f is nonnegative. Also note that $\mathcal{L}^*$ is convex, so f^* is measurable.

Step 4. Verify the feasibility and optimality.

By construction, f^* is feasible since

$$\mathcal{W}_p(f^*, \boldsymbol{\nu}) = \max_{u, v \in C_b(\Xi): \int v d\boldsymbol{\nu} = 0} \left\{ \int_\Xi u f^* dA : u(\xi) \leq \Phi_{v^*}(\xi), \ \forall \xi \in \Xi \right\} \leq \theta^p.$$

We verify that f^* is primal optimal. From the concavity of $\mathcal{L}$, we have $\mathcal{L}(f^*(\xi)) \geq p^* \mathcal{L}^*(-\frac{\partial}{\partial \lambda -} \mathcal{L}(\lambda^* \Phi_{v^*}(\xi))) + (1 - p^*) \mathcal{L}(-\frac{\partial}{\partial \lambda +} \mathcal{L}^*(\lambda^* \Phi_{v^*}(\xi)))$. Using (2.35) and the fact that

$\mathcal{L}(x) - ax = \mathcal{L}^*(a)$ for all $x \in -\partial\mathcal{L}^*(a)$, we have

$$\begin{aligned}
v_P &\geq \int_\Xi \mathcal{L}(f^*(\xi))A(d\xi) \\
&\geq p^* \int_\Xi \mathcal{L}(-\frac{\partial}{\partial\lambda-}\mathcal{L}(\lambda^*\Phi_{v^*}(\xi)))A(d\xi) + (1-p^*)\int_\Xi \mathcal{L}^*(-\frac{\partial}{\partial\lambda+}\mathcal{L}^*(\lambda^*\Phi_{v^*}(\xi)))A(d\xi) \\
&= p^* \int_\Xi \Big[\mathcal{L}^*(\lambda^*\Phi_{v^*}(\xi)) - \lambda^*\Phi_{v^*}(\xi)\frac{\partial}{\partial\lambda-}\mathcal{L}^*(\lambda^*\Phi_{v^*}(\xi))\Big]A(d\xi) \\
&\quad + (1-p^*)\int_\Xi \Big[\mathcal{L}^*(\lambda^*\Phi_{v^*}(\xi)) - \lambda^*\Phi_{v^*}(\xi)\frac{\partial}{\partial\lambda+}\mathcal{L}^*(\lambda^*\Phi_{v^*}(\xi))\Big]A(d\xi) \\
&= v_D.
\end{aligned}$$

Hence we conclude that there exists a worst-case distribution of the form (2.36). In particular, when $\mathcal{L}(\cdot) = \sqrt{\cdot}$, we have $\partial L^*(a) = \frac{1}{4a^2}$, $f^*(\xi) = \frac{1}{4\lambda^{*2}\Phi_{v^*}(\xi)^2}$, and $\lambda^* = \sqrt{\int_\Xi \frac{1}{4\theta^p\Phi_{v^*}}dA}$. We remark that we obtain a slightly more compact form than that in [89].

2.7 Concluding Remarks

In this chapter, we developed a constructive proof method to derive the dual reformulation of distributionally robust stochastic optimization with Wasserstein distance under a general setting. Such approach allows us to obtain a precise structural description of the worst-case distribution and connects the distributionally robust stochastic optimization to classical robust programming. Based on our results, we obtain many theoretical and computational implications. For the future work, extensions to multi-stage distributionally robust stochastic optimization will be explored.

CHAPTER 3
DISTRIBUTIONAL ROBUSTNESS AND REGULARIZATION IN STATISTICAL LEARNING

3.1 Overview

This chapter is based on [90]. To be consistent with statistical learning literature, in this chapter we use x, y, z to denote random variables.

Statistical learning theory [91] provides a framework for learning functional dependencies from past data, so as to make better predictions and decisions for the future. Typically, a statistical learning problem is written as

$$\min_{\beta \in \mathcal{D}} \ \mathbb{E}_{(x,y)\sim \boldsymbol{\mu}_{\rm true}}[\ell_\beta(x,y)].$$

Here the term $\ell_\beta(x,y)$ is defined as $\ell_\beta(x,y) := \ell(f(x;\beta),y)$, where the function $f(x;\beta)$ is the hypothesis function parameterized by $\beta \in \mathcal{D}$, the function ℓ is the per-sample loss function, and (x,y) is an input-output random vector with probability distribution $\boldsymbol{\mu}_{\rm true}$ on the data space $\Xi \subset \mathbb{R}^K$.

In practice, the true data-generating distribution $\boldsymbol{\mu}_{\rm true}$ might be unknown. However, a sample of observations from the distribution $\boldsymbol{\mu}_{\rm true}$ is often available. Thus, a common practice is to replace the expected risk under the unknown true distribution $\boldsymbol{\mu}_{\rm true}$ with the empirical risk under the empirical distribution $\boldsymbol{\nu}_n$, which is constructed from n data points $\{(\hat{x}^i,\hat{y}^i)\}_{i=1}^n$. Thereby we obtain the following *empirical risk minimization* problem:

$$\min_{\beta \in \mathcal{D}} \ \mathbb{E}_{(x,y)\sim \boldsymbol{\nu}_n}[\ell_\beta(x,y)].$$

Empirical risk minimization often yields solutions which perform well on the training data,

but may perform poorly on out-of-sample data. This is known as the overfitting phenomenon. A core aim of statistical learning is to design algorithms with a better generalization ability, i.e., the ability to perform well on new, previously unseen data. To reduce the generalization error, a great number of *regularization* methods have been proposed. A typical regularization problem can be represented as

$$\min_{\beta\in\mathcal{D}} \mathbb{E}_{(x,y)\sim\boldsymbol{\nu}_n}[\ell_\beta(x,y)] + \theta \cdot J(\beta), \tag{Regularization}$$

where J is the regularization penalty which may depend on ℓ and $\boldsymbol{\nu}_n$.

In this chapter, we establish a connection between DRSO with Wasserstein distance and regularization (Regularization). More precisely, we consider the following Wasserstein DRSO problem

$$\min_{\beta\in\mathcal{D}} \sup_{\boldsymbol{\mu}\in\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{(x,y)\sim\boldsymbol{\mu}}[\ell_\beta(x,y)], \tag{Wasserstein-DRSO}$$

where

$$\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n) \;:=\; \big\{\boldsymbol{\mu}\in\mathcal{P}(\Xi):\; \mathcal{W}_p(\boldsymbol{\mu},\boldsymbol{\nu}_n)\le\theta\big\}.$$

Here we explicitly express the dependence of $\mathfrak{M}$ on θ and $\boldsymbol{\nu}_n$. We establish a connection between (Wasserstein-DRSO) and the regularization (Regularization). Such connection has several important methodological and algorithmic implications.

In Section 3.3.1, we show an exact equivalence between (Wasserstein-DRSO) with $p = 1$ and norm-penalty regularization for the linear loss function class. Next in Section 3.3.2, for a broad class of smooth loss functions, we show that (Wasserstein-DRSO) is asymptotically equivalent to the following regularization problem

$$\min_{\beta\in\mathcal{D}} \mathbb{E}_{(x,y)\sim\boldsymbol{\nu}_n}[\ell_\beta(x,y)] + \theta \cdot \left\|\nabla_{(x,y)}\ell_\beta\right\|_{\boldsymbol{\nu}_n,p_*}, \tag{3.1}$$

where $p_* = \frac{p}{p-1}$ and the penalty term $\left\|\nabla_{(x,y)}\ell_\beta\right\|_{\boldsymbol{\nu}_n,p_*}$ represents the empirical p_*-norm (see

Definition 3.2 in Section 3.2) of the gradient of the loss function with respect to the data.

In Section 3.4, we provide a new interpretation of the discrete choice models from the perspective of distributional robustness, based on the equivalence result for linear optimization. Discrete choice models are used to describe decision makers' choices among a finite set of alternatives, and have attracted a lot of interest in economics, marketing, operations research and management science. Many choice models can be based on random utility theory [92, 93, 94], in which the utilities of alternatives are random, and each consumer chooses the alternative with the highest realized utility. A recent approach, called semi-parametric model [52, 95, 96] combines the idea of random utility theory and distributional robustness, in which the distribution of the random utilities is given by the worst-case distribution over a set of distributions for an expected utility maximization problem. Choice models can also be based on representative agent model [97, 98], in which a representative agent maximizes a weighted sum of utilities of alternatives plus a regularization term that encourages diversification. We refer to [99] for a study on relations between these choice models. Based on our equivalence result, the representative agent choice model can be derived from an ambiguity-averse representative agent choosing a choice probability vector that maximizes the expected utility. This connection offers a new economic intuition for the generalized extreme value choice models, which was introduced in the literature pure mathematically.

The asymptotic equivalence suggests a principled way to regularize statistical learning problems, namely, by solving the regularization problem (3.1). This is illustrated by the training of Wasserstein generative adversarial networks in Section 3.5.1 and estimation of mixed logit models 3.5.2.

Finally, the conclusion of the paper is made in Section 3.6. Auxiliary results are provided in the Appendix B.

3.2 Preliminary

We introduce several definitions and review some results on DRSO with Wasserstein distance.

Definition 3.1 (∞-Wasserstein distance)**.** The ∞-Wasserstein distance between distributions $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\Xi)$ is defined as

$$\mathcal{W}_\infty(\boldsymbol{\mu}, \boldsymbol{\nu}) := \inf_{\boldsymbol{\gamma} \in \Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})} \boldsymbol{\gamma}\text{-}\operatorname*{ess\,sup}_{\Xi \times \Xi} \mathsf{d}(z, z'),$$

where $\Gamma(\boldsymbol{\mu}, \boldsymbol{\nu})$ denotes the set of all Borel probability distributions on $\Xi \times \Xi$ with marginal distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, and $\boldsymbol{\gamma}\text{-}\operatorname*{ess\,sup}_{\Xi \times \Xi} \mathsf{d}(z, z')$ expresses the essential supremum of $\mathsf{d}(\cdot, \cdot)$ with respect to the measure $\boldsymbol{\gamma}$.

Given the empirical distribution $\boldsymbol{\nu}_n := \frac{1}{n} \sum_{i=1}^n \delta_{\hat{z}^i}$, consider

$$\min_{\beta \in \mathcal{D}} \sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{z \sim \boldsymbol{\mu}}[\ell_\beta(z)]. \tag{Wasserstein-DRSO}$$

Note that here we do not restrict to supervised learning, in which case $z = (x, y)$. Recall from Chapter 2 that problem (Wasserstein-DRSO) admits a strong duality reformulation, as shown by the following lemma.

Lemma 3.1. *Assume $h : \Xi \to \mathbb{R}$ satisfies $h(z) \leq C(\|z - z^0\| + 1)^q$ for some constants $q \in [1, \infty)$, $C > 0$, $z^0 \in \Xi$, and for all $z \in \Xi$. Then for the set $\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)$ ($p \in [q, \infty)$), it holds that*

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[h(z)] = \min_{\lambda \geq 0} \Big\{ \lambda \theta^p + \frac{1}{n} \sum_{i=1}^n \sup_{z \in \Xi} \big[h(z) - \lambda \big\| z - \hat{z}^i \big\|^p \big] \Big\},$$

and for $\mathfrak{M}_\infty^\theta(\boldsymbol{\nu}_n)$, it holds that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_\infty^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[h(z)] \;=\; \sup_{z^i\in\Xi, i=1,\ldots,n} \Big\{\frac{1}{n}\sum_{i=1}^n h(z^i) :\; \big\|z^i-\hat{z}^i\big\| \le \theta\Big\}.$$

Proof of Lemma 3.1. If $p \in [1,\infty)$, the result follows from Corollary 2 in Chapter 2. If $p=\infty$, the result follows from Theorem 3 in [100] by setting $R_0=\infty$. □

We remark that the condition $p \ge q$ in Lemma 3.1 is necessary, as otherwise the worst-case loss will be infinity.

Definition 3.2 (Empirical norm)**.** Let $\|\cdot\|$ be some norm on $\mathbb{R}^K$. The *empirical p-norm* $\|h\|_{\boldsymbol{\nu}_n,p}$ of $h:\Xi\to\mathbb{R}^K$ is defined as for $p\in[1,\infty)$,

$$\|h\|_{\boldsymbol{\nu}_n,p} \;:=\; \Big(\frac{1}{n}\sum_{i=1}^n \big\|h(\hat{z}^i)\big\|^p\Big)^{1/p},$$

and for $p=\infty$,

$$\|h\|_{\boldsymbol{\nu}_n,\infty} \;:=\; \max_{1\le i\le n} \big\|h(\hat{z}^i)\big\|.$$

3.3 Equivalence between Distributional Robustness and Regularization

In this section, we show that for many common statistical learning problems, the DRSO problem (Wasserstein-DRSO) and the regularization problem (Regularization) are closely related. In Section 3.3.1, we show an exact equivalence between the two problems for the linear function class. In Section 3.3.2, we show that for some smooth function class, the two problems are asymptotically equivalent.

3.3.1 Exact Equivalence for the Linear Function Class

In this subsection, we consider the linear function class in any of the following cases:

(i) [Regression] $\ell_\beta(z)=\ell(\beta^\top x-y)$, where $z=(x,y)\in\Xi=(\mathbb{R}^K,\|\cdot\|)\times(\mathbb{R},|\cdot|)$, and $\|(x,y)-(x',y')\|_\Xi=\|x-x'\|+|y-y'|$;

(ii) [Classification] $\ell_\beta(z) = \ell(y \cdot \beta^\top x)$, where $z = (x, y) \in \Xi = (\mathbb{R}^K, \|\cdot\|) \times (\{-1, 1\}, \mathbb{I})$, where $\mathbb{I}(u) = 0$ if $u = 0$ and $\mathbb{I}(u) = \infty$ otherwise, and $\|(x, y) - (x', y')\|_\Xi = \|x - x'\| + \mathbb{I}(y - y')$;

(iii) [Unsupervised learning] $\ell_\beta(z) = \ell(\beta^\top z)$, where $z \in \Xi = (\mathbb{R}^K, \|\cdot\|)$.

Here $\ell : \mathbb{R} \to \mathbb{R}$ is some univariate Lipschitz continuous loss function (cf. Examples 3.1-3.3 below). Note that for any Lipschitz function, Rademacher's theorem (see, e.g., Theorem 2.14 in [72]) implies that the set of differentiable points of ℓ is dense in $\mathbb{R}$. We have the following equivalence result.

Theorem 3.1 (Linear predictor)**.** *Under the setup described as above, suppose ℓ is L_ℓ-Lipschitz continuous. Let $\mathcal{T}$ be the set of points in $\mathbb{R}$ at which ℓ is differentiable. Assume either $\lim_{t \in \mathcal{T}, t \to \infty} \ell'(t) = L_\ell$ or $\lim_{t \in \mathcal{T}, t \to -\infty} \ell'(t) = -L_\ell$. Then for the regression (Case (i) above), it holds that*

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;=\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot L_\ell \cdot \max(\|\beta\|_*, 1),$$

and for the classification and the unsupervised learning (Cases (ii)(iii) above), it holds that

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;=\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot L_\ell \cdot \|\beta\|_*,$$

Remark 3.1. Comparing to Theorem 3.1(ii) and Remark 3.15 in [43], we do not need convexity assumptions on ℓ, nor do we need any assumption on the data distribution (such as non-separability for SVM as specified in [39]). Instead, we require certain conditions on the asymptotics of the loss function. by relaxing the convexity assumption on ℓ. In the case of the classification, the metric structure on Ξ indicates that there is no uncertainty in the label variable y. Such assumption holds for many applications, including many image-related tasks (e.g., ImageNet competition [101]) in which the sample images are correctly labeled. On the other hand, if there is uncertainty in the label, the equivalence no longer

holds, but the DRSO can be reduced to some convex program (see, e.g., [41, 43]).

Proof of Theorem 3.1. Using Lemma 3.1 we have that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)]-\mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] \;=\; \min_{\lambda\geq 0}\left\{\lambda\theta+\frac{1}{n}\sum_{i=1}^n \sup_{z\in\Xi}\left[\ell_\beta(z)-\ell_\beta(\hat{z}^i)-\lambda\left\|z-\hat{z}^i\right\|_\Xi\right]\right\}.$$

To unify the notation for different cases, we define

$$\tilde{\beta} := \begin{cases} (\beta,-1), & \text{Case (i)},\\ (\beta,0), & \text{Case (ii)},\\ \beta, & \text{Case (iii)}. \end{cases}$$

Since ℓ is L_ℓ-Lipschitz, for any $\hat{z}^i$ and any $z\in\Xi$ (in the case of classification, z and $\hat{z}^i$ should have identical label assignments, see Remark 3.1), it holds that

$$\ell_\beta(z)-\ell_\beta(\hat{z}^i) \;\leq\; L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}\cdot\left\|z-\hat{z}^i\right\|_\Xi,$$

where $\|\cdot\|_{\Xi^*}$ represents the dual norm to the norm on Ξ. Thus, for any $\lambda\geq L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}$,

$$\sup_{z\in\Xi}\left[\ell(\beta^\top z)-\ell(\beta^\top\hat{z}^i)-\lambda\left\|z-\hat{z}^i\right\|_\Xi\right] \;=\; 0.$$

Note that $\lambda = L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}$ is a dual feasible solution, thereby

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)]-\mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] \;\leq\; \theta\cdot L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}.$$

On the other hand, note that the Lebesgue differentiation theorem (see, e.g., Theorem 1.6.11 in [102]) implies that for any $t_0<t$, $\ell(t)-\ell(t_0)=\int_{t_0}^t \ell'(s)ds$. When $\lim_{t\in\mathcal{T},t\to\infty}\ell'(t)=L_\ell$, we obtain that for any $\hat{z}^i$ and any $\lambda<L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}$,

$$\lim_{t\to\infty}\; \ell_\beta\big(\hat{z}^i+t\tilde{\beta}/||\tilde{\beta}||_{\Xi^*}\big)-\ell_\beta(\hat{z}^i)-\lambda t=\infty.$$

Similarly, when $\lim_{t\in\mathcal{T},t\to-\infty}\ell'(t)=-L_\ell$, we obtain that for any $\hat{z}^i$ and any $\lambda<L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}$,

$$\lim_{t\to\infty}\ \ell_\beta\big(\hat{z}^i-t\tilde{\beta}/||\tilde{\beta}||_{\Xi^*}\big)-\ell_\beta(\hat{z}^i)-\lambda t=\infty.$$

Therefore we conclude that

$$\min_{\lambda\geq 0}\left\{\lambda\theta+\frac{1}{N}\sum_{i=1}^{n}\sup_{z\in\Xi}\left[\ell_\beta(z)-\ell_\beta(\hat{z}^i)-\lambda\left\|z-\hat{z}^i\right\|_\Xi\right]\right\}=\min_{\lambda\geq L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}}\{\lambda\theta\}=\theta\cdot L_\ell\cdot\left\|\tilde{\beta}\right\|_{\Xi^*}.$$

Finally, note that in Case (i),

$$\|(\beta,-1)\|_{\Xi^*}=\sup_{x,y}\left\{\beta^\top x-y:\ \|x\|+|y|\leq 1\right\}=\max(\|\beta\|_*,1),$$

which completes the proof. □

Example 3.1 (Absolute deviation regression)**.** Let $\ell(t)=|t|$. Then ℓ is 1-Lipschitz. By Theorem 3.1, we have the equivalence

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)}\ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta^\top z)]\ =\ \mathbb{E}_{\boldsymbol{\nu}_n}[\ell(\beta^\top z)]\ +\ \theta\cdot\max(\|\beta\|_*,1).$$

Example 3.2 (Classification)**.** Let $\ell(z)=\ell(y\cdot\beta^\top x)$, where ℓ is any 1-Lipschitz loss function, such as the hinge loss $\max(1-y\cdot\beta^\top x,0)$, or the logistic loss $\log(1+\exp(-y\cdot\beta^\top x))$. Using Theorem 3.1, we obtain that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)}\ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta^\top z)]\ =\ \mathbb{E}_{\boldsymbol{\nu}_n}[\ell(\beta^\top z)]\ +\ \theta\cdot\|\beta\|_*,$$

which recovers Remark 3.15 in [43].

The result in Theorem 3.1 can be easily generalized to the following case, whose proof is similar to that of Theorem 3.1 and thus omitted.

Corollary 3.1. *Let* $\Xi=(\mathbb{R}^K,\|\cdot\|)$. *Suppose there exists a positive integer* M *such that*

$\ell_\beta(z) = \max_{1\le m\le M} \ell_m({\beta^m}^\top z)$, *where* $\beta^m \in \mathbb{R}^K$, $\beta = [\beta^1; \ldots; \beta^M]$, *and* $\ell_m : \mathbb{R} \to \mathbb{R}$ *is* L_m*-Lipschitz continuous. Let* $\mathcal{T}_m$ *be the set of points in* $\mathbb{R}$ *at which* ℓ_m *is differentiable. Assume for each* $1 \le m \le M$*, either* $\lim_{t\in\mathcal{T}_m, t_m\to\infty} \ell_m'(t) = L_m$ *or* $\lim_{t\in\mathcal{T}_m, t_m\to-\infty} \ell_m'(t) = -L_m$*. Then it holds that*

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta^\top z)] \;=\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell(\beta^\top z)] \;+\; \theta \cdot \max_{1\le m\le M} L_m \left\|\beta^m\right\|_* .$$

Example 3.3 (Piecewise-linear convex loss)**.** Let $\Xi = (\mathbb{R}^K, \|\cdot\|)$. Set ℓ_m in Corollary 3.1 to be the identity function, i.e., $\ell_m(z) = z$ for all $z \in \Xi$. Then we recovers Remark 6.6 in [40]:

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell(\beta^\top z)] \;=\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell(\beta^\top z)] \;+\; \theta \cdot \max_{1\le m\le M} \left\|\beta^m\right\|_* .$$

3.3.2 Asymptotic Equivalence for the Smooth Function Class

In this subsection, we consider the class of smooth functions and present an asymptotic equivalence result between (Wasserstein-DRSO) and regularization. Assume $\ell_\beta(\cdot)$ is differentiable. With Definition 3.2 in Section 3.2, the empirical p-norm of the gradient function $\nabla_z \ell_\beta$ is expressed as

$$\left\|\nabla_z \ell_\beta\right\|_{\boldsymbol{\nu}_n,p} := \begin{cases} \left(\frac{1}{n}\sum_{i=1}^n \left\|\nabla_z \ell_\beta(z^i)\right\|_*^p\right)^{1/p}, & p \in [1,\infty), \\ \max_{1\le i\le n} \left\|\nabla_z \ell_\beta(z^i)\right\|_*, & p = \infty. \end{cases}$$

We note that the gradient is taken with respect to the data z, but not with respect to the learning parameter β, where the latter is seen much often in the machine learning literature.

Theorem 3.2 (Asymptotic equivalence)**.** *Let* $\Xi = (\mathbb{R}^K, \|\cdot\|)$*. Suppose either of the following conditions holds:*

(i) ℓ_β *is Lipschitz continuous,* $p = 1$*, and* $\boldsymbol{\mu}_{\rm true}$ *has a continuous density on* Ξ*.*

(ii) There exists a constant $\kappa \in (0, 1]$ *and a function* $h : \Xi \to \mathbb{R}$ *such that*

$$\|\nabla_z \ell_\beta(z) - \nabla_z \ell_\beta(z')\|_* \ \leq \ h(z') \cdot \|z - z'\|^\kappa, \ \ \forall z, z' \in \Xi. \tag{3.2}$$

$p \in [\kappa + 1, \infty]$ *and* $h \in L^{\frac{p}{p-\kappa-1}}(\boldsymbol{\mu}_{\mathrm{true}})$.

Let the radius sequence $\{\theta_n\}_{n=1}^\infty$ *be a sequence of positive random variables convergent to zero almost surely. Then it holds almost surely that*

$$\left| \sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^{\theta_n}(\boldsymbol{\nu}_n)} \mathbb{E}_{z\sim\boldsymbol{\mu}}[\ell_\beta(z)] - \left(\mathbb{E}_{z\sim\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta_n \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, p_*} \right) \right| = o(\theta_n),$$

where the "almost surely" is with respect to i.i.d. draws of samples from $\boldsymbol{\mu}_{\mathrm{true}}$ *and the randomness of* θ_n.

Remark 3.2. Here we allow the radius sequence $\{\theta_n\}_n$ to be random, since in practice it may be determined adaptively to the data-generating mechanism and converges to zero almost surely (with respect to i.i.d. draws of random data) as more and more data are collected.

Theorem 3.2 states that the regularization problem with penalty term $\|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, p_*}$ is a first-order approximation of (Wasserstein-DRSO), and such approximation is asymptotically exact. The gradient-norm penalty has been heuristically exploited for deep learning problems, such as adversarial training ($p = \infty$, see [35, 103]) and training of generative adversarial networks ($p = 2$, see [104]).

The proof of Theorem 3.2 is based on the following two propositions, which provides upper and lower bounds of the worst-case loss $\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{(x,y)\sim\boldsymbol{\mu}}[\ell_\beta(x, y)]$ in terms of regularization.

Proposition 3.1 (Upper bound on the worst-case loss)**.**

i) Suppose ℓ_β is L_{ℓ_β}-Lipschitz. Let $p = 1$. Then it holds that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_1^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;\le\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot L_{\ell_\beta}.$$

ii) Suppose ℓ_β is differentiable, and there exists constants $\kappa \in (0, 1]$, $q \in (1, \infty)$, $C \geq 0$ and a function $h : \Xi \to \mathbb{R}$ such that

$$\|\nabla_z \ell_\beta(z) - \nabla_z \ell_\beta(z')\|_* \le\; h(z') \cdot \|z - z'\|^\kappa + C \cdot \|z - z'\|^q, \;\; \forall z, z' \in \Xi.$$

Let $p \in [q + 1, \infty)$ if $C > 0$, and $p \in [\kappa + 1, \infty)$ if $C = 0$. Then it holds that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;\le\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, p_*} + \theta^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n, p_*} + C \cdot \theta^{q+1}.$$

Proof of Proposition 3.1. (i) follows from Proposition 6.5(i) in [40].

(ii) By the assumption on ℓ_β and the mean-value theorem, it holds that

$$\ell_\beta(z) - \ell_\beta(z') \le \|\nabla_z \ell_\beta(z')\|_* \cdot \|z - z'\| + h(z') \cdot \|z - z'\|^{\kappa+1} + C \cdot \|z - z'\|^{q+1}, \;\; \forall z, z' \in \Xi.$$

When $p = \infty$, using Lemma 3.1 yields that

$$\begin{aligned}\sup_{\boldsymbol{\mu}\in\mathfrak{M}_\infty^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] - \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] &\le \frac{1}{n}\sum_{i=1}^n \left(\left\|\nabla_z \ell_\beta(\hat{z}^i)\right\|_* \cdot \theta + h(\hat{z}^i) \cdot \theta^{\kappa+1} + C \cdot \theta^{q+1}\right) \\ &\le \theta \cdot \|\ell_\beta\|_{\boldsymbol{\nu}_n, 1} + \theta^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n, 1} + C\theta^{q+1}.\end{aligned}$$

We next consider $p \in (1, \infty)$. In view of Lemma 3.1, we provide an upper bound on

$\sup_{z\in\Xi}\big\{[\ell_\beta(z)-\ell_\beta(\hat{z}^i)]-\lambda\cdot\|z-\hat{z}^i\|^p\big\}$. We have that

$$\begin{aligned}
&\sup_{z\in\Xi}\big\{[\ell_\beta(z)-\ell_\beta(\hat{z}^i)]-\lambda\cdot\big\|z-\hat{z}^i\big\|^p\big\}\\
\leq\ &\sup_{z\in\Xi}\Big\{\big\|\nabla_z\ell_\beta(\hat{z}^i)\big\|_*\cdot\big\|z-\hat{z}^i\big\|+h(\hat{z}^i)\cdot\big\|z-\hat{z}^i\big\|^{\kappa+1}+C\cdot\big\|z-\hat{z}^i\big\|^{q+1}-\lambda\cdot\|z-z'\|^p\Big\}\\
\leq\ &\sup_{t\geq 0}\Big\{\big\|\nabla_z\ell_\beta(\hat{z}^i)\big\|_*\cdot t+h(\hat{z}^i)\cdot t^{\kappa+1}+C\cdot t^{q+1}-\lambda\cdot t^p\Big\}.
\end{aligned}$$

Using Lemma B.1 in Appendix B, for any $\delta=(\delta_1,\delta_2)>0$, it holds that

$$\begin{aligned}
&\sup_{t\geq 0}\Big\{\big\|\nabla_z\ell_\beta(\hat{z}^i)\big\|_*\cdot t+h(\hat{z}^i)\cdot t^{\kappa+1}+C\cdot t^{q+1}-\lambda\cdot t^p\Big\}\\
\leq\ &\sup_{t\geq 0}\Big\{\Big(\big\|\nabla_z\ell_\beta(\hat{z}^i)\big\|_*+\frac{p-\kappa-1}{p-1}\cdot h(\hat{z}^i)\cdot\delta_1+\frac{p-q-1}{p-1}\cdot C\cdot\delta_2\Big)\cdot t\\
&\qquad-\Big(\lambda-\frac{\kappa}{p-1}\cdot h(\hat{z}^i)\cdot\delta_1^{-\frac{p-\kappa-1}{\kappa}}-\frac{q}{p-1}\cdot C\cdot\delta_2^{-\frac{p-q-1}{q}}\Big)\cdot t^p\Big\}\\
=:\ &\sup_{t\geq 0}\big\{g_\delta(\hat{z}^i)\cdot t-(\lambda-C_\delta)\cdot t^p\big\}.
\end{aligned}$$

Solving this maximization problem over t gives

$$\sup_{t\geq 0}\big\{g_\delta(\hat{z}^i)\cdot t-(\lambda-C_\delta)\cdot t^p\big\}\ =\ \begin{cases}p^{\frac{p}{1-p}}(p-1)(\lambda-C_\delta)^{-\frac{1}{p-1}}\cdot(g_\delta(\hat{z}^i))^{\frac{p}{p-1}}, & \lambda>C_\delta,\\ +\infty, & \lambda\leq C_\delta.\end{cases}$$

It then follows Lemma 3.1 that

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)}\mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)]-\mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)]\leq\inf_{\lambda\geq C_\delta}\Big\{\lambda\theta^p+p^{\frac{p}{1-p}}(p-1)(\lambda-C_\delta)^{-\frac{1}{p-1}}\cdot\|g_\delta\|^{p_*}_{\boldsymbol{\nu}_n,p_*}\Big\}.$$

Solving the right-hand side yields

$$\sup_{\boldsymbol{\mu}\in\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)}\mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)]-\mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)]\leq\theta\cdot\|g_\delta\|_{\boldsymbol{\nu}_n,p_*}+C_\delta\cdot\theta^p.$$

Plugging in the expressions for g_δ and C_δ on the right-hand side, we have

$$\theta \cdot \|g_\delta\|_{\boldsymbol{\nu}_n,p_*} + C_\delta \cdot \theta^p = \theta \cdot \Big(\|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n,p_*} + \tfrac{p-\kappa-1}{p-1} \cdot \delta_1 \cdot \|h\|_{\boldsymbol{\nu}_n,p_*} + \tfrac{p-q-1}{p-1} \cdot C \cdot \delta_2 \Big) + \theta^p \cdot \Big(\|h\|_{\boldsymbol{\nu}_n,p_*} \cdot \tfrac{\kappa}{p-1} \cdot \delta_1^{-\frac{p-\kappa-1}{\kappa}} + C \cdot \tfrac{q}{p-1} \cdot \delta_2^{-\frac{p-q-1}{q}} \Big).$$

Minimizing over $\delta > 0$ for the right-hand side gives the result.

□

Proposition 3.2 (Lower bound on the worst-case loss)**.** *Let $\Xi = \mathbb{R}^K$. Suppose ℓ_β is differentiable, and there exists a constant $\kappa \in [0,1]$ and a function $h : \Xi \to \mathbb{R}$ such that*

$$\|\nabla_z \ell_\beta(z) - \nabla_z \ell_\beta(z')\|_* \;\le\; h(z') \cdot \|z - z'\|^\kappa, \;\; \forall z, z' \in \Xi.$$

Then for $p \in (\kappa+1, \infty]$, it holds that

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;\ge\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n,p_*} - \;\theta^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n, \frac{p}{p-\kappa-1}},$$

else for $p \in [1, \kappa+1]$, it holds that

$$\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] \;\ge\; \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n,p_*} - \;\theta^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n,\infty}.$$

Proof of Proposition 3.2. The proof uses the following observation: a lower bound on the worst-case loss is given by only considering distributions that are supported on n points. More specifically, $\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] - \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)]$ is lower bounded by the following quantity

$$\sup_{z_i \in \Xi, i=1,\ldots,n} \left\{ \frac{1}{n} \sum_{i=1}^n [\ell_\beta(z^i) - \ell_\beta(\hat{z}^i)] : \; \Big(\frac{1}{n} \sum_{i=1}^n \big\| z^i - \hat{z}^i \big\|^p \Big)^{1/p} \le \theta \right\}. \tag{3.3}$$

Indeed, for $p = \infty$, problem (3.3) is equivalent to $\sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] - \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)]$ by Lemma 3.1, and for $p \in [1, \infty)$, the feasible set of problem (3.3) can be viewed as a subset

of $\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)$, which contains distributions that are supported on at most n points. Then by the assumption on $\nabla_z \ell_\beta$ and the mean-value theorem, the quantity $\sup_{\boldsymbol{\mu}\in\mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{\boldsymbol{\mu}}[\ell_\beta(z)] - \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)]$ is lower bounded by

$$\sup_{z_i\in\Xi, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n \left[\nabla_z \ell_\beta(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\| - h(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\|^{\kappa+1} \right] : \left(\frac{1}{n}\sum_{i=1}^n \left\|z^i - \hat{z}^i\right\|^p\right)^{1/p} \leq \theta \right\},$$

which is further lower bounded by

$$\begin{aligned} &\sup_{z_i\in\Xi, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_z \ell_\beta(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\| : \left(\frac{1}{n}\sum_{i=1}^n \left\|z^i - \hat{z}^i\right\|^p\right)^{1/p} \leq \theta \right\} \\ &\qquad - \sup_{z_i\in\Xi, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n h(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\|^{\kappa+1} : \left(\frac{1}{n}\sum_{i=1}^n \left\|z^i - \hat{z}^i\right\|^p\right)^{1/p} \leq \theta \right\}. \end{aligned}$$

Note that $\Xi = \mathbb{R}^K$, then it follows from Hölder's inequality that

$$\begin{aligned} & \sup_{z_i\in\Xi, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_z \ell_\beta(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\| : \left(\frac{1}{n}\sum_{i=1}^n \left\|z^i - \hat{z}^i\right\|^p\right)^{1/p} \leq \theta \right\} \\ = & \sup_{t_i\in\mathbb{R}, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n \nabla_z \ell_\beta(\hat{z}^i) \cdot t_i : \left(\frac{1}{n}\sum_{i=1}^n t_i^p\right)^{1/p} \leq \theta \right\} \\ = & \theta \cdot \left\|\nabla_z \ell_\beta\right\|_{\boldsymbol{\nu}_n, p_*}, \end{aligned}$$

and that

$$\begin{aligned} & \sup_{z_i\in\Xi, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n h(\hat{z}^i) \cdot \left\|z^i - \hat{z}^i\right\|^{\kappa+1} : \left(\frac{1}{n}\sum_{i=1}^n \left\|z^i - \hat{z}^i\right\|^p\right)^{1/p} \leq \theta \right\} \\ = & \sup_{t_i\in\mathbb{R}, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n h(\hat{z}^i) \cdot t_i^{\kappa+1} : \left(\frac{1}{n}\sum_{i=1}^n t_i^p\right)^{1/p} \leq \theta \right\}. \end{aligned}$$

Therefore, the results follow by observing that when $p > \kappa + 1$,

$$\sup_{t_i\in\mathbb{R}, i=1,\ldots,n} \left\{ \frac{1}{n}\sum_{i=1}^n h(\hat{z}^i) \cdot t_i^{\kappa+1} : \left(\frac{1}{n}\sum_{i=1}^n t_i^p\right)^{1/p} \leq \theta \right\} = \theta^{\kappa+1} \cdot \left\|h\right\|_{\boldsymbol{\nu}_n, \frac{p}{p-\kappa-1}},$$

and when $p \leq \kappa + 1$,

$$\sup_{t_i \in \mathbb{R}, i=1,\ldots,n} \left\{ \frac{1}{n} \sum_{i=1}^{n} h(\hat{z}^i) \cdot t_i^{\kappa+1} : \ \left(\frac{1}{n} \sum_{i=1}^{n} t_i^p \right)^{1/p} \leq \theta \right\} \ \leq \ \theta^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n, \infty}.$$

□

With Propositions 3.1 and 3.2, we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. When $p > 1$, Propositions 3.1 and 3.2 imply that

$$\left| \sup_{\boldsymbol{\mu} \in \mathfrak{M}_p^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{z \sim \boldsymbol{\mu}}[\ell_\beta(z)] - \mathbb{E}_{z \sim \boldsymbol{\nu}_n}[\ell_\beta(z)] - \theta_n \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, p_*} \right| \leq \theta_n^{\kappa+1} \cdot \|h\|_{\boldsymbol{\nu}_n, \frac{p}{p-\kappa-1}}.$$

Then the integrability assumption on h and the Law of Large Numbers ensure that the remainder on the right-hand side is $o(\theta_n)$ almost surely.

When $p = 1$ and ℓ_β is Lipschitz continuous, we denote by L_{ℓ_β} the smallest Lipschitz constant of ℓ_β. In view of Propositions 3.1 and 3.2, it suffices to prove

$$\mathbb{P}\Big\{ \lim_{n \to \infty} \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, \infty} = L_{\ell_\beta} \Big\} = 1,$$

where $\mathbb{P}$ denotes that the probability is taken with respect to i.i.d. draws of samples from $\boldsymbol{\mu}_{\text{true}}$. To this end, observe from (3.2) that, for any $\epsilon > 0$,

$$\delta \ := \ \boldsymbol{\mu}_{\text{true}}\big\{ z \in \Xi : \ \|\nabla_z \ell_\beta(z)\|_* > L_{\ell_\beta} - \epsilon \big\} > 0.$$

It then follows that

$$\mathbb{P}\big\{ \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, \infty} \ < \ L_{\ell_\beta} - \epsilon \big\} \ < \ (1 - \delta)^n,$$

Thus,

$$\sum_{n=1}^{\infty} \mathbb{P}\big\{ \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, \infty} \ < \ L_{\ell_\beta} - \epsilon \big\} \ < \ \sum_{n=1}^{\infty} (1 - \delta)^n \ < \ \infty.$$

Then by Borel-Cantelli lemma (see, e.g. [105]), $\|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n,\infty}$ converges to L_{ℓ_β} almost surely, which completes the proof. □

3.4 Application of the Equivalence in Discrete Choice Modeling

In this subsection, we consider linear optimization $\ell_\beta(z) = \beta^\top z$, and apply the equivalence result to discrete choice modeling. This is a special case of the linear function class with Lipschitz loss considered in Section 3.3.1, but here we allow a more general metric structure on Ξ other than the norm $\|\cdot\|$, and allow arbitrary nominal distribution $\boldsymbol{\nu} \in \mathcal{P}(\Xi)$ rather than the empirical distribution $\boldsymbol{\nu}_n$.

For ease of exposition, we consider a soft-constrained version of (Wasserstein-DRSO):

$$\max_{\beta \in \mathcal{D}} \inf_{\boldsymbol{\mu} \in \mathcal{P}(\Xi)} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\beta^\top z] + \eta \cdot \mathcal{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}) \right\}. \tag{3.4}$$

where $\eta > 0$. Problem (3.4) can be interpreted as follows. Suppose z represents a vector of random utilities of d products. Let $\mathcal{D} = \{\beta \in \mathbb{R}^K_+ : \sum_{k=1}^K \beta_k = 1\}$ be the set of choice probability vectors for these products. Consider a consumer who wants to maximize her total expected utility, but is ambiguous about the true distribution of the random utilities. We can model such ambiguity through a Wasserstein ball of distributions centered at some reference distribution $\boldsymbol{\nu}$. Thus, problem (3.4) is interpreted as an ambiguity-averse consumer choosing a choice probability vector $\beta \in \mathcal{D}$ that maximizes the worst-case total expected utility.

We make the following assumptions on the data space (Ξ, d) in the Definition 4.2 of Wasserstein distance. Let Ξ be a linear subspace of $\mathbb{R}^K$. Assume that the distance function d is translation-invariant, i.e., it can be expressed as $\mathsf{d}(\cdot,\cdot) = \mathsf{D}(\cdot - \cdot)$ for some function $\mathsf{D} : \Xi \to \mathbb{R}$. We also assume that $\mathsf{D}(u)$ is strictly convex and even in the sense that $\mathsf{D}(u) = \mathsf{D}(|u|)$ for all $u \in \Xi$, where $|u|$ represents component-wise absolute value function. Furthermore, we assume that $\mathsf{D}(u)$ is non-decreasing on positive orthant, i.e.,

$\mathsf{D}(u_1, \ldots, u_K)$ is non-decreasing in each component on $\{u \in \Xi : u \geq 0\}$.

All the above assumptions seems to be natural if we view D as a function that describes consumer's attitude towards ambiguity. The faster D grows, the larger penalty is imposed on the deviation from the nominal distribution, and the larger penalty, the less likely of perturbations of utilities from the nominal distribution, and thus the more certain about the nominal distribution the consumer is. An extreme case is when $\mathsf{D}(0) = 0$ and $\mathsf{D}(u) = +\infty$ for all $u \neq 0$, where u represents the deviation, then the consumer is completely certain about the nominal distribution. Indeed, in this case the only distribution that makes the inner minimization problem of (3.4) finite is the nominal distribution.

Theorem 3.3 (Choice model)**.** *Set $\bar{z} = \mathbb{E}_{\boldsymbol{\nu}}[z]$. Then under the above setup, problem (3.4) is equivalent to*

$$\max_{\beta \in \mathcal{D}} \left\{ \beta^\top \bar{z} - \eta \cdot \mathsf{D}^*\left(\frac{\beta}{\eta}\right) \right\}, \tag{3.5}$$

where D^* is the convex conjugate of D. The optimal solution β^0 satisfies

$$\beta^0 = \eta \cdot \nabla \mathsf{D}(\bar{z} + c\mathbf{1}),$$

for some $c \in \mathbb{R}$ and an all-one vector $\mathbf{1} := (1, 1, \ldots, 1)^\top$.

Remark 3.3. The regularization problem (3.5) is equivalent to the formulation of representative agent choice model [97, 98], which states that the choice probability vector is given by the solution of the regularization problem (3.5). Thus, the equivalence result provides a new interpretation of the representative agent choice model from the perspective of distributional robustness, i.e., the choice probability equals the optimal solution to the distributionally robust utility maximization problem (3.4).

Theorem 3.3 states that the vector of the choice probabilities is proportional to the gradient of the distance function D at $\bar{z}+c\mathbf{1}$. The following examples show that with proper choices of D, the vector of the choice probabilities is also proportional to the gradient of the

distance function D at $\bar{z}$. As a multivariate function, the distance function D summarizes the consumer's *cross-ambiguity attitude* towards multiple products.

Example 3.4 (Multinomial logit)**.** Let $\mathsf{D}_{MNL}(u) = \sum_{k=1}^{d} \exp(u_k)$. We have

$$\nabla_k \mathsf{D}_{MNL}(\bar{z} + c\mathbf{1}) = \exp(\bar{z}_k + c) = \exp(\bar{z}_k) \cdot \exp(c).$$

Thus β_k^0 is proportional to $\exp(\bar{z}_k)$. Using the normalization condition $\beta \in \Delta$, we recover the multinomial logit choice model

$$\beta_k^0 = \frac{\exp(\bar{z}_k)}{\sum_{j=1}^{K} \exp(\bar{z}_j)}.$$

For the multinomial logit model, D_{MNL} is separable and is additive in each deviation u_k, which suggests that the consumer is cross-ambiguity neutral for all products, i.e., the consumer's judgment on the likely of perturbations of utilities are independent across products.

Example 3.5 (Nested logit)**.** Let $\mathcal{G}$ be a partition of $\{1, \ldots, d\}$. Each element of the partition is called a nest. Denote by u_g the sub-vector of u whose indices belong to $g \in \mathcal{G}$, and denote by $g(k)$ the nest that k belongs to. Let

$$\mathsf{D}_{NL}^{\mathcal{G}}(u) = \sum_{g \in \mathcal{G}} \|\exp(u_g)\|_{1/\tau_g} = \sum_{g \in \mathcal{G}} \Big(\sum_{k \in g} \exp(|u_k|/\tau_g) \Big)^{\tau_g},$$

where $\tau_g > 0$ are parameters. Then

$$\begin{aligned}
\beta_k^0 \;&\propto\; \nabla_k \mathsf{D}_{NL}^{\mathcal{G}}(\bar{z} + c\mathbf{1}) \\
&=\; \Big(\sum_{k':g(k')=g(k)} \exp((\bar{z}_{k'} + c)/\tau_{g(k)}) \Big)^{\tau_{g(k)}} \cdot \frac{\exp((\bar{z}_k + c)/\tau_{g(k)})}{\sum_{k':g(k')=g(k)} \exp((\bar{z}_{k'} + c)/\tau_{g(k)})} \\
&=\; \exp(\bar{z}_k/\tau_{g(k)}) \Big(\sum_{k':g(k')=g(k)} \exp(\bar{z}_{k'}/\tau_{g(k)}) \Big)^{\tau_{g(k)}-1},
\end{aligned}$$

where $\propto$ represents "proportional to". With proper scaling, we recover the nested logit

model (see, e.g., (4.2) in [106]). Unlike the multinomial logit model, the distance function $\mathsf{D}_{NL}^{\mathcal{G}}$ for the nested logit model is not additively separable in products, but only additively separable in nests. This indicates that the consumer's judgment on the likely of perturbations of utilities are interrelated across the products within the same nest, and are independent across different nests.

Example 3.6 (GEV)**.** Let $\mathsf{D}_{GEV}(u) = \mathsf{D}_0(\exp(u_1), \ldots, \exp(u_d))$ for some strictly convex differentiable function $\mathsf{D}_0 : \mathbb{R}_+^d \to \mathbb{R}$. Assume that D_0 is homogeneous, i.e., $\mathsf{D}_0(tY_1, \ldots, tY_d) = t^s \mathsf{D}_0(Y_1, \ldots, Y_d)$ for all $t > 0$ and some $s > 0$. It follows that

$$\begin{aligned}
\beta_k^0 \;&\propto\; \nabla_k \mathsf{D}_{GEV}(\bar{z} + c\mathbf{1}) \\
&=\; \exp(c) \cdot \exp(\bar{z}_k) \cdot \nabla_k \mathsf{D}_0((\exp(c) \cdot \exp(\bar{z}_1), \ldots, \exp(c) \cdot \exp(\bar{z}_d))) \\
&\propto\; \exp(\bar{z}_k) \cdot \nabla_k \mathsf{D}_0(\exp(\bar{z}_1), \ldots, \exp(\bar{z}_d)).
\end{aligned}$$

This exactly corresponds to the expression of Generalized Extreme Value (GEV) choice models proposed by [93]. We note that as pointed out in the original GEV framework, the function D_0 has little economic intuition. Our equivalence result Theorem 3.3 endows the function D_0 with an economic interpretation – it reflects the consumer's ambiguity attitude.

Proof of Theorem 3.3. Using Definition 4.2 of Wasserstein distance, we have that

$$\inf_{\boldsymbol{\mu} \in \mathcal{P}(\Xi)} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\beta^\top z] + \eta \cdot \mathcal{W}_1(\boldsymbol{\mu}, \boldsymbol{\nu}) \right\} = \inf_{\boldsymbol{\gamma} \in \Gamma(\mathbb{P}, \boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\gamma}}[\beta^\top z + \eta \cdot \mathsf{D}(z - z')].$$

For a random vector (z, z') with joint distribution $\boldsymbol{\gamma}$, we denote by $\boldsymbol{\gamma}_{z'}$ the condition distribution of z given z'. It then follows from the disintegration theorem (see, e.g., Theorem 5.3.1 in [107]) that

$$\inf_{\boldsymbol{\gamma} \in \Gamma(\mathbb{P}, \boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\gamma}}[\beta^\top z + \eta \cdot \mathsf{D}(z - z')] = \inf_{\{\boldsymbol{\gamma}_{z'}\}_{z' \subset \mathcal{P}(\Xi)}} \int_\Xi \int_\Xi [\beta^\top z + \eta \cdot \mathsf{D}(z - z')] \, \boldsymbol{\gamma}_{z'}(dz) \, \boldsymbol{\nu}(dz').$$

Using interchangeability principle (see, e.g, Theorem 7.80 in [1]), we exchange the infi-

mum and integration:

$$\begin{aligned}&\inf_{\{\boldsymbol{\gamma}_{z'}\}_{z'}\subset\mathcal{P}(\Xi)}\int_{\Xi}\int_{\Xi}[\beta^{\top}z+\eta\cdot\mathsf{D}(z-z')]\;\boldsymbol{\gamma}_{z'}(dz)\;\boldsymbol{\nu}(dz')\\=&\int_{\Xi}\inf_{\boldsymbol{\gamma}_{z'}\in\mathcal{P}(\Xi)}\left\{\int_{\Xi}[\beta^{\top}z+\eta\cdot\mathsf{D}(z-z')]\boldsymbol{\gamma}_{z'}(dz)\right\}\boldsymbol{\nu}(dz').\end{aligned}$$

Observe that $\inf_{\boldsymbol{\gamma}_{z'}\in\mathcal{P}(\Xi)}\left\{\int_{\Xi}[\beta^{\top}z+\eta\cdot\mathsf{D}(z-z')]\boldsymbol{\gamma}_{z'}(dz)\right\}$ is attained on some Dirac measure, thereby

$$\begin{aligned}\inf_{\boldsymbol{\gamma}_{z'}\in\mathcal{P}(\Xi)}\left\{\int_{\Xi}[\beta^{\top}z+\eta\cdot\mathsf{D}(z-z')]\boldsymbol{\gamma}_{z'}(dz)\right\}&=\inf_{z\in\Xi}\;[\beta^{\top}z+\eta\cdot\mathsf{D}(z-z')]\\&=\beta^{\top}z'+\inf_{z\in\Xi}\;[\beta^{\top}(z-z')+\eta\cdot\mathsf{D}(z-z')].\end{aligned}$$

Combining the equations above yields

$$\begin{aligned}&\inf_{\boldsymbol{\mu}\in\mathcal{P}(\Xi)}\left\{\mathbb{E}_{\boldsymbol{\mu}}[\beta^{\top}z]+\eta\cdot\mathcal{W}_1(\boldsymbol{\mu},\boldsymbol{\nu})\right\}\\=&\mathbb{E}_{\boldsymbol{\nu}}[\beta^{\top}z']+\int_{\Xi}\inf_{z\in\Xi}[\beta^{\top}(z-z')-\eta\cdot\mathsf{D}(z-z')]\;\boldsymbol{\nu}(dz').\end{aligned}$$

By the assumption that Ξ is a linear space, we have that for any $z'\in\Xi$,

$$\inf_{z\in\Xi}[\beta^{\top}(z-z')-\eta\cdot\mathsf{D}(z-z')]=\inf_{u\in\Xi}[\beta^{\top}u-\eta\cdot\mathsf{D}(u)].$$

Thus it follows that

$$\inf_{\boldsymbol{\mu}\in\mathcal{P}(\Xi)}\left\{\mathbb{E}_{\boldsymbol{\mu}}[\beta^{\top}z]+\eta\cdot\mathcal{W}_1(\boldsymbol{\mu},\boldsymbol{\nu})\right\}=\beta^{\top}\bar{z}-\eta\cdot\sup_{u\in\Xi}\;[u^{\top}(\beta/\eta)-\mathsf{D}(-u)].$$

Observe that the strict convexity of D implies that the optimal solution of $\sup_{u\in\Xi}[u^{\top}(\beta/\eta)-\mathsf{D}(-u)]$ is unique. Moreover, the non-negativity of β and $\mathsf{D}(u)=\mathsf{D}(|u|)$ for all u imply

that the optimal solution is non-negative, and

$$\sup_{u\in\Xi}[u^\top(\beta/\eta)-\mathsf{D}(-u)]=\sup_{u\in\Xi}[u^\top(\beta/\eta)-\mathsf{D}(|u|)]=\sup_{u\in\Xi}[u^\top(\beta/\eta)-\mathsf{D}(u)].$$

Hence problem (3.4) is equivalent to

$$\max_{\beta\in\mathcal{D}}\left\{\beta^\top\bar{z}-\eta\cdot\sup_{u\in\Xi}\,[u^\top(\beta/\eta)-\mathsf{D}(u)]\right\}=\max_{\beta\in\mathcal{D}}\{\beta^\top\bar{z}-\eta\cdot\mathsf{D}^*(\beta/\eta)\}. \tag{3.6}$$

We next drive the optimality condition of (3.6). Consider the following relaxation of (3.6):

$$\max_{\beta\in\mathbb{R}^K}\left\{\beta^\top\bar{z}-\eta\cdot\mathsf{D}^*(\beta/\eta):\ \sum_{k=1}^d\beta_k=1\right\}.$$

Let β^0 be its optimal solution, which is unique since D^* is also strictly convex. Its first-order optimality condition yields that the optimal solution β^0 should satisfy

$$\bar{z}+c\mathbf{1}=\eta\cdot\nabla\mathsf{D}^*(\beta^0/\eta)/\eta=\nabla\mathsf{D}^*(\beta^0/\eta),$$

where $\nabla\mathsf{D}^*$ represents the gradient of D^*, c is the Lagrangian multiplier of the constraint $\sum_{k=1}^d\beta_k=1$, and $\mathbf{1}=(1,\ldots,1)^\top$. Since $\nabla\mathsf{D}^*=(\nabla\mathsf{D})^{-1}$ (see, e.g., Exercise 3.40 in [77]), it follows that

$$\beta^0=\eta\cdot\nabla\mathsf{D}(\bar{z}+c\mathbf{1}).$$

Observe that $\nabla\mathsf{D}$ is non-negative by our assumption, therefore β^0 is also an optimal solution of (3.6), which completes the proof. □

3.5 A Principled Way to Regularize Learning Problems

In this section, we discuss the algorithmic implications of the equivalence between distributionally robust framework (Wasserstein-DRSO) and regularization. We have the following two observations.

1) The worst-case loss, which is obtained by solving a minimax problem (Wasserstein-DRSO), admits an asymptotically exact approximation that can be obtained by solving a single minimization problem

$$\min_{\beta\in\mathcal{D}} \mathbb{E}_{\boldsymbol{\nu}_n}[\ell_\beta(z)] + \theta_n \cdot \|\nabla_z \ell_\beta\|_{\boldsymbol{\nu}_n, p_*}, \tag{3.7}$$

which is often much easier to solve.

2) The distributionally robust formulation (Wasserstein-DRSO) suggests a principled way to regularize statistical learning problem — by solving the gradient-norm regularization problem (3.7).

To illustrate the second point, we consider the *adversarial examples*, which has received much attention recently in adversarial learning. [108] pointed out that several classification models in machine learning, including state-of-the-art neural networks, are not robust to adversarial examples at all. The adversarial examples are obtained from a slightly perturbation of a correctly classified training example. [35] suggested a fast way of generating adversarial examples by the following perturbation:

$$\hat{x}_k^i \;\mapsto\; \hat{x}_k^i \;+\; \theta \cdot \mathrm{sgn}\left(\partial_{x_k}\left(\ell_\beta(\hat{x}^i, \hat{y}^i)\right)\right), \quad \forall 1 \le k \le d,\ \forall 1 \le i \le n, \tag{3.8}$$

where sgn is the sign function, and its argument $\partial_{x_k}\left(\ell_\beta(\hat{x}^i, \hat{y}^i)\right)$ represents the partial derivative of the loss function with respect to the k-th component of the i-th data point. For example, for logistic regression $\ell_\beta(x, y) = \ell(-y \cdot (\beta^\top x))$, where $\ell(\cdot) = \log(1 + \exp(\cdot))$, data point $\hat{x}^i$ is perturbed to

$$\hat{x}^i + \theta \cdot \mathrm{sgn}(\beta). \tag{3.9}$$

To improve the robustness of logistic regression, [35] proposed to solve the following ad-

versarial version of logistic regression

$$\min_{\beta\in\mathcal{D}} \ \ell(-y\cdot(\beta^\top x+b-\theta\left\|\beta\right\|_1)).$$

We now show that how one can use distributionally robust formulation (Wasserstein-DRSO) and its regularization approximation (Regularization) to find an algorithm that is robust to adversarial examples. Let us consider a special case of (Wasserstein-DRSO), in which $p=\infty$ and $\|(x,y)-(x',y')\| = \left\|x-x'\right\|_\infty + \mathbb{I}(y-y')$, where $\mathbb{I}(u)=0$ if $u=0$ and $\mathbb{I}(u)=\infty$ if $u\neq 0$. By Lemma 3.1, the dual problem of (Wasserstein-DRSO) is given by

$$\min_{\beta\in\mathcal{D}} \sup_{(x^i,\hat{y}^i)\in\Xi, i=1,\ldots,n} \Big\{\frac{1}{n}\sum_{i=1}^n \ell_\beta(x^i,\hat{y}^i): \ \left\|x_k^i-\hat{x}_k^i\right\|_\infty \leq \theta, \ \forall 1\leq i\leq n, \ 1\leq k\leq d\Big\}. \tag{3.10}$$

This can be viewed as the most *adversarial* way to construct the Wasserstein ball, since each component of each data point can be perturbed arbitrarily within a distance θ. The corresponding regularization approximation (Regularization) can be rewritten as

$$\min_{\beta\in\mathcal{D}} \ \mathbb{E}_{\nu_n}[\ell_\beta(x,,y)] + \theta\cdot\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^d \mathrm{sgn}(\partial_{x_k}\ell_\beta(\hat{x}^i,\hat{y}^i)), \tag{3.11}$$

where sgn is the sign function. Observe that the direction $\mathrm{sgn}(\partial_{x_k}\ell_\beta(\hat{x}^i,\hat{y}^i))$ in (3.11) is exactly the adversarial direction along which data point is perturbed as described in (3.8). Then simply by solving (3.11), we would find solutions that are robust against adversarial examples. In particular, for logistic regression, (3.11) is equivalent to the distributionally robust formulation (3.10), and (3.9) describes the worst-case distribution of the distributionally robust logistic regression.

Next, we apply the regularization scheme (Regularization) to two important classes of problems: the training of Generative Adversarial Networks (GANs) in deep learning and the estimation of mixed logit choice models.

3.5.1 Training of the Wasserstein Generative Adversarial Networks in Deep Learning

We start with a brief introduction to Generative Adversarial Networks (GANs) [109, 110]. GANs are a powerful class of *generative models*, which aim to answer the following central question in unsupervised learning:

How to learn a probability distribution from data?

The primary motivations to study this question includes: (i) learning conditional distributions that are used in reinforcement learning and semi-supervised learning, (ii) generating realistic samples for real-world tasks such as single image super-resolution [111], art creation [112], and image-to-image translation [113], and (iii) testing our ability to represent and manipulate multi-modal high-dimensional probability distributions.

To answer the question above, the classical approach is to perform density estimation. This is often done by considering a parametric family of distributions and find one that maximizes the likelihood function on the data. However, this approach does not work well for high-dimensional data-generating distributions in many applications [66], such as natural images, symbols in natural language corpora, and audio waveform containing speech. Instead of estimating the explicit density function, an implicit approach works as follows. Let Z be a random variable with a simple and fixed distribution $\mathbb{P}_0$ (such as a Gaussian distribution). Passing the random variable through a parametric function g_β (called a generator and usually described by a neural network), we denote the distribution of the random variable $g_\beta(Z)$ by $\boldsymbol{\mu}_\beta$, which is called the model distribution. By varying β in the parameter space Θ, we can find one model distribution that is "closest" to the empirical distribution. GANs are well known examples of this approach. Among lots of variants of GANs that exploit different notion of closeness between distributions, *Wasserstein GAN (WGAN)* [114] has recently attracts great attention in deep learning, as its training requires few parameter tuning, which is ideal for many deep learning problems. In WGAN, the closeness between the model distribution μ_β and the empirical distribution $\boldsymbol{\nu}_n$ is measured

by the 1-Wasserstein distance:

$$\min_{\beta \in \Theta} \mathcal{W}_1(\boldsymbol{\mu}_\beta, \boldsymbol{\nu}_n). \tag{3.12}$$

Estimation of the Wasserstein distance between high-dimensional distributions is hard. In fact, the sample complexity exponentially depends on the dimension [115]. In WGAN, the following method is used to estimate $\mathcal{W}_1(\boldsymbol{\mu}_\beta, \boldsymbol{\nu}_n)$. By Kantorovich-Rubinstein duality [70], the 1-Wasserstein distance can be written as

$$\sup_f \ \mathbb{E}_{x \sim \boldsymbol{\nu}_n}[f(x)] - \mathbb{E}_{z \sim \boldsymbol{\mu}_\beta}[f(z)], \tag{3.13}$$

where the inner supremum is taken over the class of 1-Lipschitz functions (or L-Lipschitz functions with any $L > 0$):

$$|f(x) - f(z)| \leq \|x - z\|, \quad \forall x \in \operatorname{supp} \boldsymbol{\nu}_n, \ z \in \operatorname{supp} \boldsymbol{\mu}_\beta. \tag{3.14}$$

To compute (3.13), the set of Lipschitz functions is often parametrized through a critic neural network $\{f_w\}_{w \in W}$, as the gradient computation for a neural network is efficient. The conceptual diagram of WGAN is shown in Figure 3.1. Samples from the standard Gaussian distribution are fed into the generator network g_β, whose outputs (fake images with distribution P_β) are compared with the true samples (real images with empirical distribution $\boldsymbol{\nu}_n$). The comparison is done by approximately computing $\mathcal{W}_1(\boldsymbol{\mu}_\beta, \boldsymbol{\nu}_n)$ using another neural network f_w.

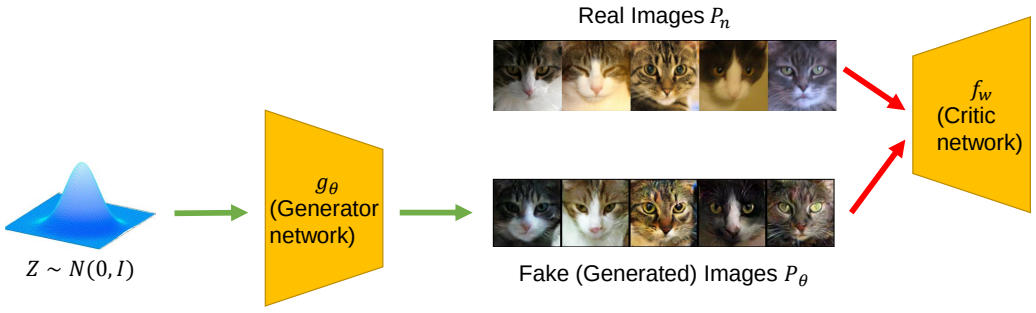


Figure 3.1: Conceptual model of WGAN

To enforce the Lipschitz condition (3.14) on the neural network f_w, [114] proposes to clip the weight vector w within a rectangle

$$\{w : -c \leq w \leq c\},$$

and formulating a minimax problem

$$\min_{\beta\in\Theta} \max_{-c\leq w\leq c} \mathop{\mathbb{E}}_{x\sim\boldsymbol{\nu}_n}[f_w(x)] - \mathop{\mathbb{E}}_{z\sim\boldsymbol{\mu}_\beta}[f_w(z)]. \tag{3.15}$$

However, the weight clipping does not describe the set of 1-Lipschitz functions, but only a subset of L-Lipschitz functions, where L depends on c. Yet another natural way to enforce the Lipschitz condition is considering a soft-constraint penalty term

$$\min_{\beta\in\Theta} \max_{w\in W} \mathop{\mathbb{E}}_{x\sim\boldsymbol{\nu}_n}[f_w(x)] - \mathop{\mathbb{E}}_{z\sim\boldsymbol{\mu}_\beta}[f_w(z)] - \lambda \cdot \mathop{\mathbb{E}}_{x\sim\boldsymbol{\nu}_n, z\sim\boldsymbol{\mu}_\beta}\Big[\Big(\frac{|f(x)-f(z)|}{\|x-z\|} - 1\Big)_+^2\Big]. \tag{3.16}$$

We here propose a new training method using ideas from Section 3.3.2. We consider a distributional robust formulation of problem (3.12) for computing $\mathcal{W}_1(\boldsymbol{\nu}_n, \boldsymbol{\mu}_\beta)$ by regularizing the inner maximization of problem (3.15):

$$\max_{w\in W} \min_{\boldsymbol{\mu}\in\mathfrak{M}_2^\theta(\boldsymbol{\nu}_n)} \mathbb{E}_{x\sim\boldsymbol{\mu}}[f_w(x)] - \mathbb{E}_{z\sim\boldsymbol{\mu}_\beta}[f_w(z)].$$

Using Theorem 3.2, this can be approximated by the regularization problem

$$\max_{w\in W} \mathbb{E}_{x\sim\boldsymbol{\nu}_n}[f_w(x)] - \mathbb{E}_{z\sim\boldsymbol{\mu}_\beta}[f_w(z)] - \theta \cdot \|\nabla_x f_w(x)\|_{\boldsymbol{\nu}_n,2}.$$

Then we add a soft-constraint penalty term and as a result, we formulate a distributionally

robust WGAN (DR-WGAN) objective:

$$\min_{\beta\in\Theta}\max_{w\in W}\mathop{\mathbb{E}}_{x\sim\boldsymbol{\nu}_n}[f_w(x)]-\mathop{\mathbb{E}}_{z\sim\boldsymbol{\mu}_\beta}[f_w(z)]-\lambda\cdot\mathop{\mathbb{E}}_{x\sim\boldsymbol{\nu}_n,z\sim\boldsymbol{\mu}_\beta}\Big[\Big(\frac{|f(x)-f(z)|}{\|x-z\|}-1\Big)_+^2\Big]-\theta\cdot\|\nabla f_w(x)\|_{\boldsymbol{\nu}_n,2}\,. \tag{DR-WGAN}$$

We compare our proposed approach with two above-mentioned benchmarks: the weight clipping approach [114] and the soft-constraint without regularization (3.16). The neural networks architecture is similar to the setup in [114], in which both the generator and critic are 4-layer ReLU-MLP with 512 hidden units. Each approach is trained using stochastic gradient descent, in which the learning rate is adjusted using Adam algorithm [116] with default choice of parameters. The detailed training algorithm for DR-WGAN is presented in Algorithm 2, which is modification of the algorithm in [114].

Algorithm 2 The proposed DR-WGAN.

while β has not converged **do**
  for $t = 0, \ldots, n_{\text{critic}}$ **do**
    for $i = 1, \ldots, m$ **do**
      Sample $x \sim \boldsymbol{\nu}_n$, $z \sim \mathbb{P}_0$
      $L^{(i)} \leftarrow f_w(x) - f_w(g_\beta(z)) + \lambda \cdot \Big(\frac{|f(x)-f(z)|}{\|x-z\|} - 1\Big)_+^2 + \theta \cdot \|\nabla_w f_w(x)\|_2$
    end for
    $w \leftarrow \text{Adam}\big(\nabla_w(\frac{1}{m}\sum_{i=1}^m L^{(i)}), w\big)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^m \sim \mathbb{P}_0$ a batch of prior samples.
  $\beta \leftarrow \text{Adam}\big(-\nabla_\beta(\frac{1}{m}\sum_{i=1}^m f_w(g_\beta(z^{(i)}))), \beta\big)$
end while

We test the performance on two standard datasets: CIFAR-10 data set [117] and the CAT dataset [118]. The CIFAR-10 dataset includes 60,000 32×32 color images in 10 classes, with 6000 images per class. The performance is measured by the inception score [119] which mimics the human eye's judgment on the similarity between the generated images and the real images, and the higher inception score, the better the performance is. Figure 3.2 plots the inception scores over the course of training with WGAN and our proposed DR-WGAN. We observe that our method converges much faster and achieves a higher inspection score. We do not plot the training curve for the soft-constraint approach

(3.16), as it does not even converge and the inception score remains at a relatively low value.

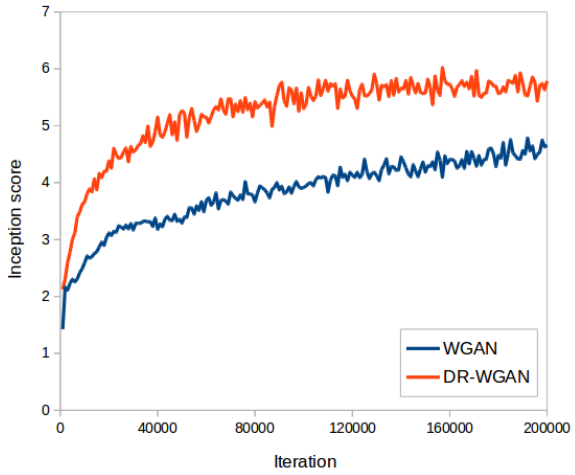


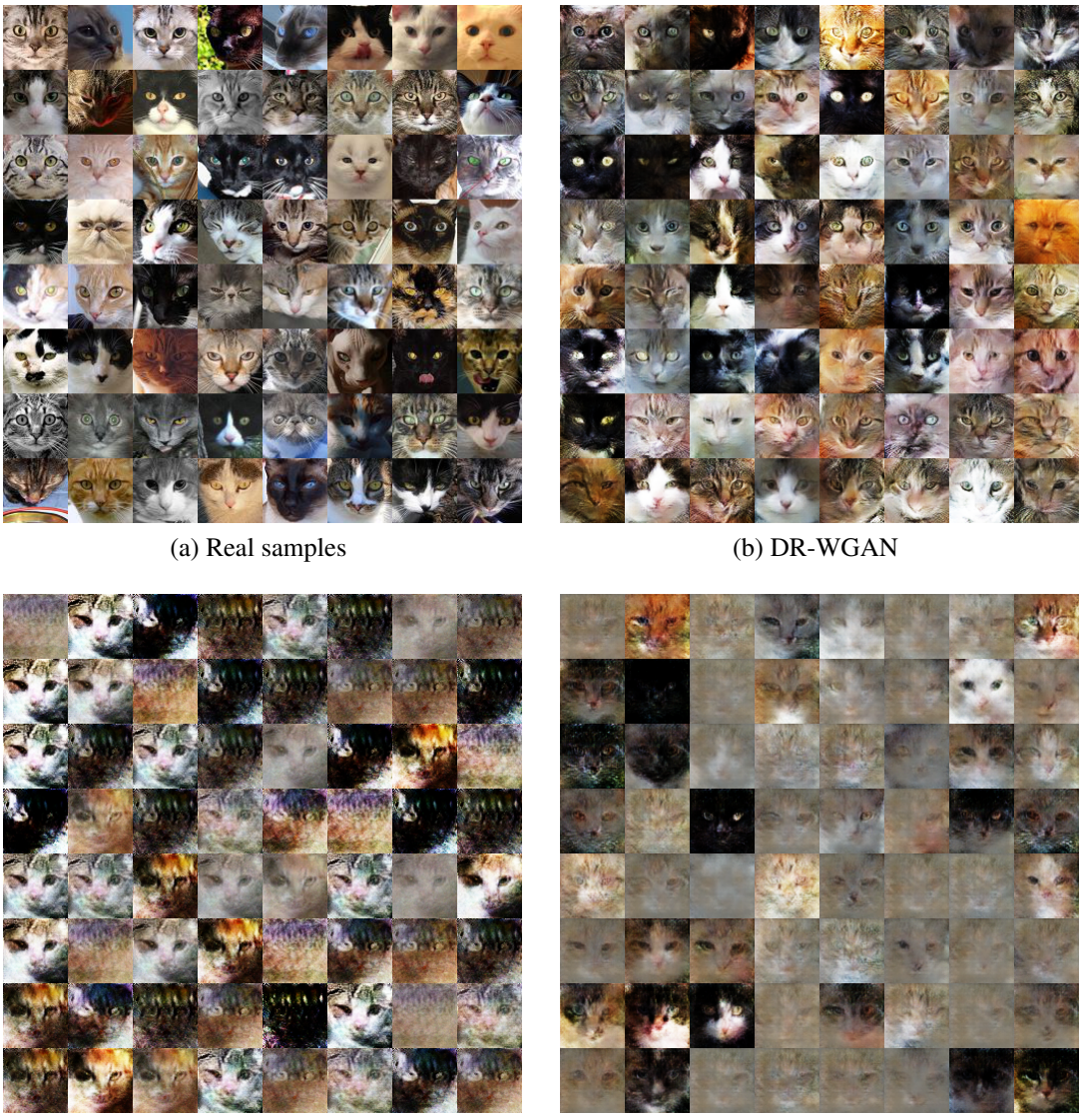
Figure 3.2: Inception scores for CIFAR-10 over generator iterations

The CAT dataset consists of 10,000 cat images, which are preprocessed such that cat faces are aligned, and scaled to 64×64 [120]. Figure 3.3a plots the real images sampled from the dataset. For each of the three approaches, we generate images from the learned parametric distribution with different random seeds, i.e., we input the generator network g_β with i.i.d Gaussian samples, and the generated images are shown in Figure 3.3b-3.3d. We observe that the image generated by WGAN exhibits mode collapse, i.e., a lack of variety, and comparing to the other two benchmarks, DR-WGAN generates images that are much more close to reality.

3.5.2 Learning Heterogeneous Customers' Preferences with Mixed Logit Model

In this subsection, we propose a new regularization scheme for the maximum likelihood estimation (MLE) of the mixed logit choice model that learns heterogeneous customers' preferences on different alternatives.

The problem originates from an airline revenue management project collaborated with a major airline, and the readers are referred to [121] for a detailed description. In this



(a) Real samples

(b) DR-WGAN

(c) WGAN (weight clipping)

(d) Soft Lipschitz-constraint penalty

Figure 3.3: CAT dataset: real and generated samples

project, an important task is to learn airline customers' choice behavior, i.e, predicting the probability that a customer chooses an airline ticket from a set of alternatives. The market that we are working with contains more than 30 parallel nonstop flights per day between two major hubs, with departure times ranging from 7 am to 10 pm. This market is featured by that most of the customers are business passengers, who are highly time-sensitive and usually only consider fairly narrow departure time windows.

In [121], a mixed logit model is used to capture customers' taste heterogeneity towards

different departure times as well as predicting the choice probabilities. We here provide a brief introduction to this model, and we refer the reader to Section 5 of [121] for a more detailed description and [106] for a more general introduction to the mixed logit choice model. We index the customers by $\{1, 2, \ldots, n\}$. Let A_i be the set of alternatives offered to the customer i. Each alternative $k \in A_i$ is described through a vector $x^{ik} \in \mathbb{R}^K$ of attributes including ticket price, booking time, departure time, ticket change fee, mileage gain, carrier, etc (see Table 1 in [121] for a complete description). The regression coefficients of the attributes are represented by a vector $[\beta; \tilde{\beta}]$, where the subvector β is deterministic, whereas the subvector $\tilde{\beta}$ is random with probability distribution $\pi_{\tilde{\beta}}$, which captures the taste heterogeneity among the population. The probability that the customer i chooses alternative k from the set of alternatives A_i is given by

$$\mathbb{E}_{\pi_{\tilde{\beta}}}\left[\frac{\exp([\beta; \tilde{\beta}]^\top x^{ik})}{\sum_{k' \in A_i} \exp([\beta; \tilde{\beta}]^\top x^{ik'})}\right]. \tag{3.17}$$

In this project, $\tilde{\beta}$ represents the departure time coefficients, and β represents the coefficients for all the other features. More precisely, the range of the departure time is partitioned into hourly time windows indexed by $k = 0, 1, \ldots, 14$, where index 0 corresponds to the reference category, hence $\tilde{\beta}$ is a random vector on $\mathbb{R}^{14}$. We assume $\tilde{\beta}$ is normally distributed with mean vector β and covariance matrix $\Sigma = CC^\top$, where C represents the Cholesky matrix. Thus, $\beta \in \mathbb{R}^{14}$ and $C \in \mathbb{R}^{14} \times \mathbb{R}^{14}$. The choice probability (3.17) can be written as

$$\mathbb{E}_{\xi \sim N(0,I)}\left[\frac{\exp([\beta; \beta + C\xi]^\top x^{ik})}{\sum_{k' \in A_i} \exp([\beta; \beta + C\xi]^\top x^{ik'})}\right],$$

which can be evaluated using Monte Carlo method.

Let $y_{ik} = 1$ if consumer i chooses alternative $k \in A_i$ in the observed data and let $y_{ik} = 0$ otherwise. Then the parameters β and C can be estimated using the maximum

likelihood method:

$$\min_{\beta,\beta,C} \; -\frac{1}{n}\sum_{i=1}^{n} y_{ik} \cdot \log\left(\mathbb{E}_{\xi\sim N(0,I)}\left[\frac{\exp([\beta;\beta+C\xi]^\top x^{ik})}{\sum_{k'\in A_i}\exp([\beta;\beta+C\xi]^\top x^{ik'})}\right]\right). \tag{3.18}$$

To the best of our knowledge, existing literature does not consider the regularization of problem (3.18). Based on the result in Section 3.3.2, we propose the following penalty to regularize the maximum likelihood problem (3.18):

$$\left\| y_{ik} \cdot \nabla_{x^{ik}} \log\left(\mathbb{E}_{\xi\sim N(0,I)}\left[\frac{\exp([\beta;\beta+C\xi]^\top x^{ik})}{\sum_{k'\in A_i}\exp([\beta;\beta+C\xi]^\top x^{ik'})}\right]\right)\right\|_{\boldsymbol{\nu}_n,2}. \tag{3.19}$$

The parameters $[\beta;\beta;C]$ are estimated using airline transaction-level booking data. We refer the reader to Section 4 of [121] for a complete description of airlines dataset. We use the first half year of 2012 data to train the model, and use the second half year of 2012 data to test the out-of-sample performance. We compare two approaches, maximum likelihood estimation (3.18) with and without regularization (3.19). In both approaches, the parameters are learned using mini-batch stochastic gradient with batch size 128, and the learning rate is adjusted by Adam algorithm [116] with default parameters. For a fair comparison, common random number are used in two approaches for generating initial points (from a standard Gaussian distribution), subsampling from population, and generating Monte Carlo samples to evaluate the mixed logit choice probability. The radius of the Wasserstein ball is determined by cross-validation.

Among the estimated coefficients, we are particularly interested in Σ, as it reflects the taste heterogeneity towards departure time windows. Hence, we here only focus on the solution quality of Σ for the two approaches. We plot in Figure 3.4 the heat maps of the correlation matrices $D^{-1}\Sigma D^{-1}$ estimated from the two approaches, where $D = \sqrt{\mathrm{diag}(\Sigma)}$. Each grid corresponds one element in the correlation matrix, ranging from 7 am to 9 pm. Blue, yellow and red represent -1, 0, 1, respectively.

For MLE with regularization, the correlation matrix (the heat map on the right) in-

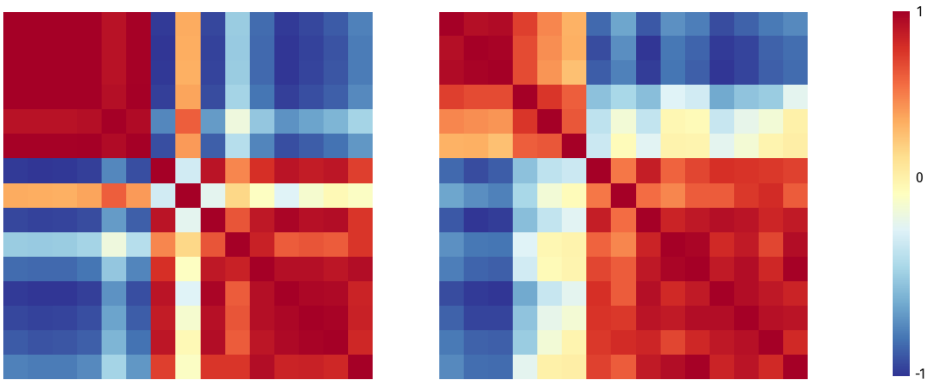


Figure 3.4: Heat maps of the correlation matrix of the taste coefficients of departure time windows. (Left: MLE without regularization. Right: MLE with regularization.)

dicates positively correlated preferences among morning/noon flight, and among afternoon/evening flight, and a negatively correlated preferences between morning/noon and afternoon/evening flights. Moreover, the correlation of preferences gradually decreases as the gap between departure time windows increases. This is consistent to our intuition. On the other hand, for MLE without regularization (the heat map on the left), although we observe similar preferences pattern, it is hard to observe a gradually change in the correlation. In particular, there is a sudden change in the sixth and seventh column. Namely, for morning/noon flights, the correlation matrix indicates the preferences are almost perfectly positively correlated, and their correlation with flights of departure time window 1:00-2:00 pm (the sixth column) suddenly changes to negative, and then becomes positive again with flights of departure time window 2:00-3:00 pm (the seventh column). Such preferences pattern seems to be hard to explain. Therefore we believe that the estimation result using MLE with regularization seems to be more plausible.

3.6 Concluding Remarks

In this chapter, we propose the Wasserstein distributionally robust formulation for solving statistical learning problems with guaranteed small generalization error. We show that it is asymptotic equivalent to a specific gradient-norm regularization problem. Such connection

provides new interpretations for regularization, and offers new algorithmic insights. For the future work, it is interesting to provide generalization error bounds for statistical learning problems based on distributional robustness, and apply the regularization scheme to other machine learning problems.

CHAPTER 4
DISTRIBUTIONALLY ROBUST OPTIMIZATION WITH KNOWN MARGINALS

4.1 Overview

This chapter is based on [122]. In Section 4.2, we motivate the study by pointing out some issues of two existing approaches regarding distributions with given marginals. In Section 4.3, we review some results on copula theory, and demonstrate how to construct copula in data-driven problems, then we describe how to use Wasserstein distance to describe the similarity between dependence structures of distributions.

Next in Section 4.4, we derive a tractable dual reformulation of problem (1.4). This generalizes the duality results in [123] and [124], in which only the marginal constraints are considered, and also generalizes the results in [24] and [62], in which only the Wasserstein constraint is considered. Our proof technique combines ideas from a refined constructive approach developed in Chapter 2, a new variational argument, and the theory of multi-dimensional Monge-Kantorovich optimal transport problem [125, 126].

For a data-driven problem in which the nominal model is the empirical copula, we show that when the objective function ℓ is a piecewise-linear convex function of the random variables, with properly chosen Wasserstein distance, the size of the convex program reformulation of the inner maximization of (1.4) only linearly depends on the dimension of the random variable, even though the support of the worst-case distribution can contain exponentially many points (Corollary 4.1). This greatly improves the scalability of our approach.

Finally, we test the performance of our formulation on two problems. The first is a mean-CVaR portfolio selection problem (Section 4.5.1), whose parameters are calibrated using real data. The numerical results show superior performance of our approach in high

dimension, as opposed to sample average approximation and distributionally robust formulation with only Wasserstein constraints. The second is nonparametric copula density estimation (Section 4.5.2). Our formulation suggests a novel estimation method. Numerical result on a real dataset illustrates promising results of our approach when the sample size is much less the dimension of the parameters.

4.2 Motivation

Our first example show that the worst-case approach described in (1.2) does not consider any information on the joint distribution at all, and thus conceivably, its worst-case distribution often involves fully correlated (i.e., comonotonic or counter-monotonic) components, which may be too extreme for many practical applications.

Example 4.1 (Over-conservative worst-case copula)**.** Consider the life insurance model described in [127, 128]. Each individual risk $\boldsymbol{\xi}_k$ has a two point distribution with $\mathbb{P}(\boldsymbol{\xi}_k = 0) = p_k$ and $\mathbb{P}(\boldsymbol{\xi}_k = \alpha_k) = 1 - p_k$, where α_k represents the value of the k-th claim, p_k denotes the survival probability of the k-th individual, and $p_1 \leq \cdots \leq p_K$. Suppose the function $\ell_x(\cdot) := \ell(x, \cdot)$ is supermodular[1], for example, the stop-loss $\max\{0, \sum_{k=1}^K \boldsymbol{\xi}_k - t\}$ of aggregate risks $\sum_{k=1}^K \boldsymbol{\xi}_k$ for some $t > 0$. The worst-case copula of (1.2) is comonotonic[2], and implies that for the corresponding worst-case distribution $\boldsymbol{\mu}^*$, it holds that

$$\mathbb{P}_{\boldsymbol{\mu}^*}[\boldsymbol{\xi}_{k+1} = 0 | \boldsymbol{\xi}_k = 0] \;\; = \;\; 1, \;\; k = 1, \ldots, K-1,$$

[1] A function is supermodular, if

$$\begin{aligned} &\ell_x(\xi_1, \ldots, \xi_k, \ldots, \xi_{k'}, \ldots, \xi_K) + \ell_x(\xi_1, \ldots, \xi_k + \epsilon, \ldots, \xi_{k'} + \delta, \ldots, \xi_K) \\ &\geq \ell_x(\xi_1, \ldots, \xi_k + \epsilon, \ldots, \xi_{k'}, \ldots, \xi_K) + \ell_x(\xi_1, \ldots, \xi_k, \ldots, \xi_{k'} + \delta, \ldots, \xi_K) \end{aligned}$$

for all $\xi \in \Xi$, $1 \leq k < k' \leq K$ and $\epsilon, \delta > 0$

[2] A distribution is comonotonic if its cumulative distribution function satisfies

$$F^{\boldsymbol{\mu}^*}(\xi_1, \ldots, \xi_K) \;\; = \;\; \min_{1 \leq k \leq K} F_k^{\boldsymbol{\mu}^*}(\xi_k), \quad \forall\, \xi.$$

which means that the death of an individual implies the deaths of all individuals with smaller survival probabilities. In particular, when $p_1 = \cdots = p_K$, $\boldsymbol{\mu}^*$ has only two possible scenarios: either all individuals are alive or they all die. Unless the insurance is for some catastrophe, this worst-case distribution seems to be unrealistic, since the dependence of mortality rates among individuals cannot be so strong.

Our second example shows that in a data-driven setting, DRSO with divergence measures (1.3) has limitations.

Example 4.2 (KL divergence ball is not suitable for data-driven problem)**.** Consider the nominal distribution is given by $N = 30$ i.i.d. observations from a Gaussian distribution. Suppose that we use this empirical distribution as the nominal distribution, then the KL divergence ball $\{\boldsymbol{\mu} : (\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^0) \leq \theta\}$ only contains distributions whose support is a subset of the nominal distribution, as indicated by the left image in Fig. 4.1. However, observe that with probability one, any two data points do not have identical coordinates in either dimension. Hence, if we also consider constraints on the marginals (1.3), then with probability one, the KL ball is a singleton containing only the empirical distribution itself. To avoid this pathological behavior, one possible remedy is to partition the space into a finite number of bins, such that each bin consists of sufficiently many empirical points. Nevertheless, there is no general guidance on how to make the partition, and it is problematic for high dimensional problems, when the number of data points is less than the dimension of the random variables.

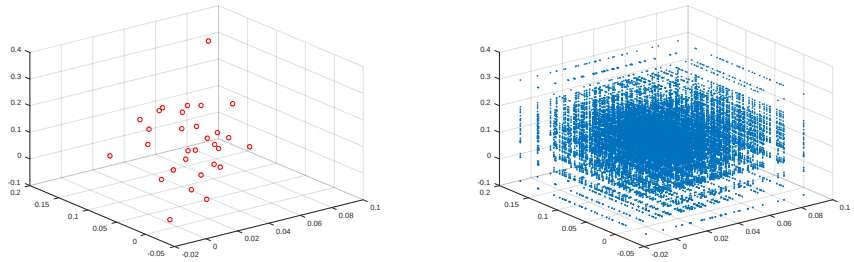


Figure 4.1: Supports of distributions within a KL divergence ball (1.3) and a Wasserstein ball (1.4)

From the modeling point of view, the advantages of using Wasserstein distance are two-fold.

(i) For copulas of distributions with highly correlated components, Wasserstein distance yields a more intuitive quantitative relationship [100], as illustrated by the following example.

Example 4.3. Table 4.1 shows various distances between copulas of Gaussian distributions $\boldsymbol{\mu}_1 = Normal(0, [1, 0.5; 0.5, 1])$, $\boldsymbol{\mu}_2 = Normal(0, [1, 0.9; 0.9, 1])$, and $\boldsymbol{\mu}_3 = Normal(0, [1, 0.999; 0.999, 1])$.

Table 4.1: Distances between copulas of Gaussian distributions

Distances	Fisher-Rao	KL	Burg entropy	Hellinger	Bhattacharya	TV	2-Wasserstein
$\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}_1}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}_2}$	1.15	1.21	0.42	0.37	0.15	0.85	0.11
$\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}_3}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}_2}$	3.26	1.81	47.20	0.75	0.81	3.71	0.07

Intuitively, distance between $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ should be smaller since both $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ are close to a comonotonic distribution. Among the distances above, only Wasserstein metric is consistent with our intuition.

(ii) When the nominal copula is an independent copula, Wasserstein distance defines a new measure of dependence, and is closely related to Spearman's ranking correlation coefficient, as indicated from Section 4.3 below.

4.3 Copulas and Wasserstein Distance between Copulas

In this section, we introduce the Wasserstein distance between copulas, and investigate its properties. In the introduction, we have mentioned that the copula is unique for a multivariate continuous distribution. However, in many data-driven problems, the nominal distribution is often finite-supported, which raises the question on the non-uniqueness of copula. To resolve this issue, we consider a slightly general notion called *subcopula*. For ease of

exposition, we do not distinguish a probability distribution and its cumulative distribution function as its meaning should be clear from the context. For example, for a distribution $\boldsymbol{\mathcal{C}}$ on $[0,1]^K$, $\boldsymbol{\mathcal{C}}(u)$ is equivalent to $\boldsymbol{\mathcal{C}}([0,u_1]\times\cdots\times[0,u_K])$. Recall that the *support* of a distribution $\boldsymbol{\mu}$ is the complement of the largest open set which has $\boldsymbol{\mu}$-measure zero.

Definition 4.1 (Subcopula and Copula)**.** A K-dimensional *subcopula* $\boldsymbol{\mathcal{C}}$ is a joint distribution with the following properties:

(i) For all $1 \le k \le K$, the k-th marginal distribution of $\boldsymbol{\mathcal{C}}$, denoted by $\boldsymbol{\mathcal{C}}_k$, has support $\operatorname{supp} \boldsymbol{\mathcal{C}}_k \subset [0,1]$.

(ii) $\boldsymbol{\mathcal{C}}_k(u) = u$ for all $u \in \operatorname{supp} \boldsymbol{\mathcal{C}}_k$.

A K-dimensional subcopula $\boldsymbol{\mathcal{C}}$ is called a K-dimensional *copula* if $\operatorname{supp} \boldsymbol{\mathcal{C}}_k = [0,1]$ for all $1 \le k \le K$.

We next restate Sklar's theorem in terms of subcopula.

Theorem 4.1 (Sklar's Theorem.)**.** *Let $\boldsymbol{\mu}$ be a K-dimensional distribution on Ξ with marginal distribution functions $F_1,\ldots,F_K$. Then there exists a unique K-subcopula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ such that for all $\xi \in \Xi$,*

$$\boldsymbol{\mu}(\xi_1,\ldots,\xi_K) = \boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}\big(F_1(\xi_1),\cdots,F_K(\xi_K)\big),$$

and $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ is a copula if the F_k's are all continuous. Conversely, for any subcopula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and marginal distribution functions $F_1,\ldots,F_K$, the equation above defines a K-dimensional distribution $\boldsymbol{\mu}$ with marginal distributions $F_1,\ldots,F_K$.

Sklar's theorem indicates that the dependence structure of a multivariate distribution is fully characterized by a unique subcopula, which becomes a copula if the marginal distributions are continuous. If we denote the inverse cumulative distribution function of each marginals by F_k^{-1}, then $\boldsymbol{\mathcal{C}}$ can be computed through the formula

$$\boldsymbol{\mathcal{C}}(u_1,\ldots,u_K) = \boldsymbol{\mu}\big(F_1^{-1}(u_1),\ldots,F_K^{-1}(u_k)\big).$$

We here list some commonly used subcopulas and copulas.

Example 4.4 (Empirical copula)**.** Let $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}_{\widehat{\xi}^i}$ be an empirical distribution, and $\hat{F}_k^{-1}$ be the inverse cumulative empirical distribution of the k-th marginal. The empirical copula [129, 130] is defined by

$$\hat{\boldsymbol{\mathcal{C}}}(u) := \frac{1}{n}\sum_{i=1}^{n}\prod_{k=1}^{K}\mathbb{1}\{\widehat{\xi}_i^k \leq \hat{F}_k^{-1}(u_k)\}.$$

Thus, empirical copula can be viewed as the empirical distribution of the rank transformed data. Note that empirical copula is a subcopula but not a copula, since $\operatorname{supp}\boldsymbol{\mathcal{C}}_k \subset \{\frac{i}{n} : 1 \leq i \leq N\}$.

Example 4.5 (Independent, comonotonic, and counter-monotonic copulas)**.**

- If $\boldsymbol{\xi}$ has mutually independent components, then it has copula $\boldsymbol{\mathcal{C}}(u) = \prod_{k=1}^{K} u_k$.
- If $\boldsymbol{\xi}$ has comonotonic components, i.e., $\boldsymbol{\xi} = (F_1^{-1}(U), \ldots, F_K^{-1}(U))$ for some distribution functions $\{F_k\}_{k=1}^{K}$ and a uniformly distributed random variable U on $[0,1]$, then $\boldsymbol{\mathcal{C}}(u) = \min_{1\leq k\leq K} u_k$.
- If $K = 2$ and $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ are counter-monotonic, i.e., $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = (F_1^{-1}(U), F_2^{-1}(1-U))$ for some distribution functions F_1, F_2 and a uniformly distributed random variable U on $[0,1]$, then $\boldsymbol{\mathcal{C}}(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$.

We next illustrate on how to construct a subcopula using the dataset described in the introduction.

Example 4.6 (Construction of an Empirical Copula)**.** The joint data of number of minutes delay for two flights on the days of week that they both operate are:

$$(30,4),(-1,0),(-5,7),(12,13),(10,0),(-5,20),(0,15),(32,58),(15,85),(30,45),$$
$$(26,30),(6,23),(40,55),(3,40),(0,-8),(11,12),(7,13),(-5,9),(-11,6),(-10,-20).$$

The additional marginal data of number of minutes delay for the more frequent flight are:

$$20, 4, 5, 48, -30, -10, -22, -3, 80, -23, 0, 26, 10, 90, 90, 24, 30, 45,$$
$$17, 35, -10, -1, 30, 5, 18, 0, 40, 16, 6.$$

We denote the joint data by $\{(\widehat{\xi_1^i}, \widehat{\xi_2^i})\}_{i=1}^n$, and the extra marginal data by $\{\widehat{\xi_1^{N+j}}\}_{j=1}^m$. The empirical copula is constructed in two steps. In the first step, we use all the marginal data information, i.e., $\{\widehat{\xi_1^i}\}_{i=1}^{N+M}$ and $\{\widehat{\xi_2^i}\}_{i=1}^n$ to estimate the marginal distributions $F_1(\xi_1)$ and $F_2(\xi_2)$. For example, we can simply use empirical cumulative distribution function, or a linear interpolation of the empirical cumulative distribution function. Using the estimated marginal distribution functions, the original joint data set is converted to $(\hat{u}_1^i, \hat{u}_2^i) = (F_1(\widehat{\xi_1^i}), F_2(\widehat{\xi_2^i}))$, $i = 1, \ldots, N$. Then in the second setup, we estimate the copula density function $c(u_1, u_2)$ using the converted joint dataset $\{(\hat{u}_1^i, \hat{u}_2^i)\}_{i=1}^n$. The scatter plots of the empirical distribution and empirical copula are shown in Figure 4.2.

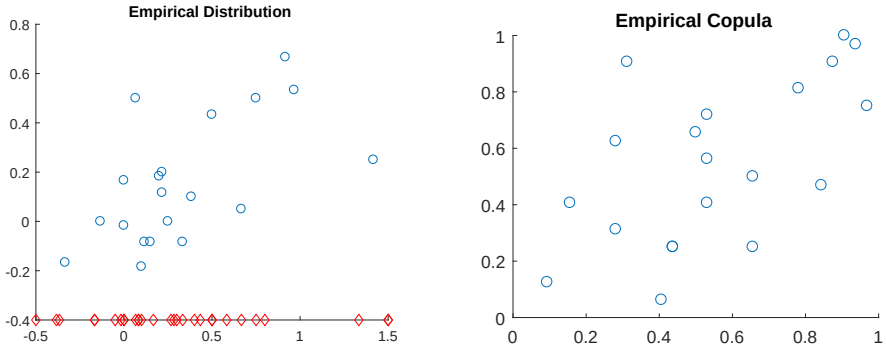


Figure 4.2: Scatter plots of empirical joint and marginal distributions and empirical copula

Let d be a metric on $[0, 1]^K$. In the case of empirical copula, d can be viewed as the distance between two relative rankings. The Wasserstein distance between two subcopulas $\boldsymbol{C}, \boldsymbol{C}^0$ is defined as follows.

Definition 4.2 (Wasserstein distance between Copulas)**.** Let $p \in [1, \infty)$. The p-Wasserstein

distance $\mathcal{W}_p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0)$ between $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0 \in \mathcal{P}([0,1]^K)$ (under metric d) is defined by

$$\mathcal{W}_p^p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0) := \min_{\boldsymbol{\gamma} \in \mathcal{P}([0,1]^{2K})} \left\{ \int_{[0,1]^{2K}} \mathsf{d}^p(u,v) \boldsymbol{\gamma}(du, dv) : \ \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0 \right\}. \quad (4.1)$$

Thus, Wasserstein distance between $\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0$ is the minimum cost (in terms of d^p) of re-distributing mass from $\boldsymbol{\mathcal{C}}$ to $\boldsymbol{\mathcal{C}}^0$. Wasserstein distance is a natural way of comparing two distributions when one is obtained from the other by perturbations.

The expression (4.1) is written in terms of the integration on $[0,1]^K$. With changing of variables, it can be equivalently represented using integration on the data space Ξ. Let $\boldsymbol{\mu}, \boldsymbol{\nu}$ be two distributions with the same marginals $\{F_k\}_k$, and denote their copulas by $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ and $\boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}}$. We define

$$\mathsf{d}_F(\xi, \zeta) := \liminf_{d(\xi^m, \xi), d(\zeta^m, \zeta) \overset{m \to \infty}{\longrightarrow} 0} \mathsf{d}\big((F_1(\xi_1^m), \ldots, F_K(\xi_K^m)), \ (F_1(\zeta_1^m), \ldots, F_K(\zeta_K^m)) \big).$$

It follows that d_F is lower semi-continuous, and d_F is a premetric [131], i.e., $\mathsf{d}_F \geq 0$ and $\mathsf{d}_F(\xi, \xi) = 0$. With these definitions, $\mathcal{W}_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}})$ can be equivalently represented as

$$\mathcal{W}_p^p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \boldsymbol{\mathcal{C}}^{\boldsymbol{\nu}}) = \min_{\boldsymbol{\gamma} \in \mathcal{P}(\Xi \times \Xi)} \left\{ \int_{\Xi^2} \mathsf{d}_F^p(\xi, \zeta) \boldsymbol{\gamma}(d\xi, d\zeta) : \ \boldsymbol{\gamma} \text{ has marginals } \boldsymbol{\mu}, \boldsymbol{\nu} \right\}.$$

Now let us consider the case when the nominal copula $\boldsymbol{\mathcal{C}}^0$ is the independent subcopula Π, which corresponds to the case where only marginal data are available. In this case, the Wasserstein distance $\mathcal{W}_p(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \Pi)$ measures the deviation of $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$ away from an independent distribution, and thus can be viewed as a measure of dependence of random variables with joint distribution $\boldsymbol{\mu}$. In particular, when $K = 2$ and $\Pi(u) = u_1 u_2$, with a special choice of d, $\mathcal{W}_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}, \Pi)$ reduces to Schweizer and Wolffs L^1-based measure of dependence [132], defined as $\int_0^1 \int_0^1 |\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(u_1, u_2) - u_1 u_2| du_1 du_2$.

Proposition 4.1. *Suppose*

$$\mathsf{d}\big((u_1,u_2),(v_1,v_2)\big) = \begin{cases} |u_1-v_1|, & if\ u_2=v_2, \\ +\infty, & o.w. \end{cases}, \text{ or } \begin{cases} |u_2-v_2|, & if\ u_1=v_1, \\ +\infty, & o.w. \end{cases}$$

Let $K=2$. Then for any distribution $\boldsymbol{\mu}$ with copula $\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}$, it holds that

$$W_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}},\Pi) = \int_0^1\int_0^1 |\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}}(u_1,u_2)-u_1u_2|du_1du_2.$$

We remark that Schweizer and Wolffs' measure of dependence is closely related to Spearman's rank correlation coefficient, which can be written as $\int_0^1\int_0^1 \big(\boldsymbol{\mathcal{C}}(u_1,u_2)-u_1u_2\big)du_1du_2$. If we set d to be the ℓ_1-norm, then $\mathcal{W}_1(\boldsymbol{\mathcal{C}}^{\boldsymbol{\mu}},\boldsymbol{\mathcal{C}}^0)$ defines a new measure of dependence which satisfies Rényi's axioms on measure of dependence [133, 132].

Proposition 4.2. *Suppose*

$$\mathsf{d}(u,v) = ||u-v||_1, \quad u,v\in[0,1]^2.$$

Let $(\boldsymbol{\xi},\boldsymbol{\zeta})$ be two random variables with continuous distribution $\boldsymbol{\mu}\in\mathcal{P}([0,1]^K)$, define

$$\omega(\boldsymbol{\xi},\boldsymbol{\zeta}) \ :=\ 12\cdot W_1(\boldsymbol{\mu},\Pi).$$

Then $\omega(\boldsymbol{\xi},\boldsymbol{\zeta})$ defines a measure of dependence that satisfies Rényi's axioms:

(i) $\omega(\boldsymbol{\xi},\boldsymbol{\zeta}) = \omega(\boldsymbol{\zeta},\boldsymbol{\xi})$.

(ii) $0\le\omega(\boldsymbol{\xi},\boldsymbol{\zeta})\le 1$.

(iii) $\omega(\boldsymbol{\xi},\boldsymbol{\zeta}) = 0$ *if and only if* $\boldsymbol{\xi}$ *and* $\boldsymbol{\zeta}$ *are independent.*

(iv) $\omega(\boldsymbol{\xi},\boldsymbol{\zeta}) = 1$ *if and only if each of* $\boldsymbol{\xi}$ *is a.s. a strictly monotone function of the other.*

(v) If f and g are strictly monotone a.s. on $\operatorname{Ran}\boldsymbol{\xi}$ *and* $\operatorname{Ran}\boldsymbol{\zeta}$ *respectively, then* $\omega(f(\boldsymbol{\xi}), g(\boldsymbol{\zeta})) = \omega(\boldsymbol{\xi}, \boldsymbol{\zeta})$.

(vi) If the joint distribution of $\boldsymbol{\xi}$ *and* $\boldsymbol{\zeta}$ *is bivariate normal with correlation coefficient* θ*, then* $\omega(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *is a strictly increasing function of* $|\theta|$.

(vii) If $(\boldsymbol{\xi}, \boldsymbol{\zeta})$ *and* $(\boldsymbol{\xi}^m, \boldsymbol{\zeta}^m)$, $m = 1, 2, \cdots$, *are pairs of random variables with joint distribution* $\boldsymbol{\mu}$ *and* $\boldsymbol{\mu}^m$ *respectively, and if the sequence* $\boldsymbol{\mu}^m$ *converges weakly to* $\boldsymbol{\mu}$*, then* $\lim_{m\to\infty} \omega(\boldsymbol{\xi}, \boldsymbol{\mu}) = \omega(\boldsymbol{\xi}, \boldsymbol{\mu})$.

4.4 Dual reformulation

In this section, we derive a dual reformulation for the inner maximization of problem (1.4). For ease of notation, we suppress variable β of ℓ. Set

$$v_P \;:=\; \sup_{\boldsymbol{\mathcal{C}}\in\mathfrak{C}} \left\{ \mathbb{E}_{\boldsymbol{\mu}}[\ell(\boldsymbol{\xi})] :\; \boldsymbol{\mu} \text{ has marginals } \{F_k\}_{k=1}^K \text{ and copula } \boldsymbol{\mathcal{C}},\; \mathcal{W}_p(\boldsymbol{\mathcal{C}}, \boldsymbol{\mathcal{C}}^0) \le \theta \right\}. \tag{4.2}$$

We assume ℓ is upper semicontinuous on Ξ, and satisfies the growth condition $\sup_{\xi\in\Xi} \frac{\ell(\xi)}{\mathsf{d}_F^p(\xi,\zeta_0)} < \infty$ for some $\zeta_0 \in \Xi$. Set $\mathfrak{M}$ to be the set of distributions $\boldsymbol{\mu}$ that is feasible to (4.2). Our main result is the following strong duality theorem.

Theorem 4.2 (Strong duality)**.** *Let* $\boldsymbol{\nu}$ *be a distribution with marginals* $\{F_k\}_{k=1}^K$ *and copula* $\boldsymbol{\mathcal{C}}^0$*. Let* Ξ_k *be the projection of* Ξ *onto the* k*-th marginal component. Then problem* (4.2) *has a strong dual problem*

$$v_D \;:=\; \inf_{\substack{\lambda\ge 0\\ f_k\in B(\Xi_k)}} \left\{ \lambda\theta^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) + \int_{\Xi} \sup_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big] \boldsymbol{\nu}(d\zeta) \right\}.$$

Before diving into the proof, we outline the proof idea as follows. To start with, it is straightforward to establish the weak duality using Lagrangian and properties of marginal distribution (Lemma 4.1). However, the difficulties in proving strong duality lie in the

non-compactness of the data space Ξ, and the semi-infinite marginal and Wasserstein constraints. To obtain the strong duality, we first assume certain compactness and continuity assumptions. Under such assumptions, we show the existence of a dual minimizer using convexification trick (see, e.g., [124, 126]) in the theory of multi-marginal optimal transport (Lemma 4.2). Next, we derive the first-order optimality condition at the dual minimizer, which helps to construct a primal optimal solution (Lemma 4.3). Finally using some limiting argument, we relax the continuity and the compactness assumption and thus complete the proof of Theorem 4.2. We only provide the proof of Lemma 4.3 here, and proofs of other lemmas and measurability of the integrand involved in the dual program are presented in the Technical Appendix.

Lemma 4.1 (Weak duality)**.** $v_P \leq v_D$.

Lemma 4.2 (Existence of dual minimizer)**.** *Assume that Ξ is compact and ℓ and d_F are Lipschitz continuous on Ξ. Then there exists a dual minimizer.*

Lemma 4.3 (Strong duality under compactness and continuity assumption)**.** *Assume that Ξ is compact and ℓ and d_F are Lipschitz continuous on Ξ. Then $v_P = v_D$.*

Proof of Lemma 4.3. We start with establishing the first-order optimality condition of the dual problem. We perform a variational analysis on the dual objective function at $(\lambda^*, \{f_k^*\}_k)$. For each $1 \leq k \leq K$, let $\{g_{km}\}_{m=1}^{\infty}$ be a Schauder basis of $B(\Xi_k)$. For any $n \in \mathbb{Z}_+$, we define a function

$$\Phi_n(\lambda, \epsilon, \zeta) := \sup_{\xi \in \Xi} \Big\{ \ell(\xi) - \sum_k f_k^*(\xi_k) - \sum_k \sum_{m=1}^{n} \epsilon_{km} g_{km}(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta) \Big\}. \tag{4.3}$$

By Lemma C.1 in Appendix, Φ is random lower semi-continuous. Moreover, for all $\zeta \in \Xi$,

$\Phi(\cdot,\cdot,\cdot,\zeta)$ is a convex function on $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{nK}$. We further define

$$h_n(\lambda,\epsilon) = \lambda\theta^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t)F_k(dt) + \sum_{k=1}^K \sum_{m=1}^n \int_{\Xi_k} \epsilon_{km} g_{km}(t) F_k(dt)$$
$$+ \int_\Xi \sup_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^K f_k^*(\xi_k) - \sum_{k=1}^K \sum_{m=1}^n \epsilon_{km} g_{km}(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big]\boldsymbol{\nu}(d\zeta).$$

Then by generalized Moreau-Rockafellar theorem (see, e.g., Theorem 7.47 in [134]), for any $(\lambda,\epsilon) \in \operatorname{dom} h_n$ it holds that

$$\partial h_n(\lambda,\epsilon) = \left(\theta, \Big[\int_{\Xi_k} g_{km}(t)F_k(dt)\Big]_{\substack{1\le k\le K\\ 1\le m\le n}}\right)^\top - \int_\Xi \partial_{\lambda,\epsilon}\Phi_n(\lambda,\epsilon,\zeta)\boldsymbol{\nu}(d\zeta) + \mathcal{N}(\lambda,\epsilon),$$

where $\mathcal{N}(\lambda,\epsilon)$ stands for the normal cone at (λ,ϵ) to the feasible region $\mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}^{nK}$. Furthermore, it follows from Theorem 2.4.18 in [135] that

$$\partial_{\lambda,\epsilon}\Phi(\lambda,\epsilon,\zeta) = \operatorname{conv}\Big\{\Big(\mathsf{d}_F^p(F(\xi(\zeta)),F(\zeta)), [g_{km}(\xi_k(\zeta))]_{\substack{1\le k\le K\\ 1\le m\le n}}\Big)^\top :$$
$$\xi(\zeta) \in \arg\max_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^K f_k^*(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big]\Big\}.$$

Set

$$T(\zeta) := \arg\max_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^K f_k^*(\xi_k) - \lambda^* \mathsf{d}_F^p(\xi,\zeta)\Big].$$

The first-order optimality condition $0 \in \partial h_n(\lambda^*, 0)$ implies that there exists $0 \le r^* \le \theta$ with $\lambda^*(\theta - r^*) = 0$, such that

$$\begin{aligned} &\left(r^*, \Big[\int_{\Xi_k} g_{km}(t)F_k(dt)\Big]_{\substack{1\le k\le K\\ 1\le m\le n}}\right) \\ &\in \int_\Xi \operatorname{conv}\Big\{\Big(\mathsf{d}^p(\xi(\zeta),\zeta), [g_{km}(\xi_k(\zeta))]_{\substack{1\le k\le K\\ 1\le m\le n}}\Big) : \xi(\zeta) \in T(\zeta)\Big\}\boldsymbol{\nu}(d\zeta). \end{aligned} \tag{4.4}$$

We construct a primal optimal solution. (4.4) suggests that there is a measurable selection $z(\zeta)$ of $\operatorname{conv}\{(\mathsf{d}^p(\xi(\zeta),\zeta), [g_{km}(\xi_k(\zeta))]_{k,m}), \xi(\zeta) \in T(\zeta)\}$, such that $\int_\Xi z(\zeta)\boldsymbol{\nu}(d\zeta) =$

$(r^*, [\int_{\Xi_k} g_{km}(t)F_k(dt)]_{k,m})$. Each $z(\zeta)$ can be represented as

$$z(\zeta) = \mathbb{E}_{\boldsymbol{\gamma}_\zeta^n}\left[\left(\mathsf{d}^p(\xi(\zeta), \zeta), [g_{km}(\xi_k(\zeta))]_{1\le k\le K, 1\le m\le n}\right)\right],$$

for some finite probability distribution $\boldsymbol{\gamma}_\zeta^n \in \mathcal{P}(T(\zeta))$, and the measurability of $z(\zeta)$ implies the measurability of $\boldsymbol{\gamma}_\zeta^n$ (as a function of ζ). Thus, there exists a probability kernel $\{\boldsymbol{\gamma}_\zeta^n\}_{\zeta\in\Xi}$ such that each $\boldsymbol{\gamma}_\zeta^n$ is a probability distribution on $T(\zeta)$ and satisfies

$$\begin{aligned} r^* &\le \theta, \\ \lambda^*(\theta - r^*) &= 0, \\ \int_{\Xi^2} g_{km}(\xi_k(\zeta))\boldsymbol{\gamma}_\zeta^n(d\xi)\boldsymbol{\nu}(d\zeta) &= \int_{\Xi_k} g_{km}(t)F_k(dt),\ \forall 1\le k\le K, 1\le m\le n. \end{aligned} \tag{4.5}$$

Now define a probability measure $\boldsymbol{\mu}^n$ by

$$\boldsymbol{\mu}^n(A) := \int_\Xi \boldsymbol{\gamma}_\zeta^n(A)\boldsymbol{\nu}(d\zeta),\ \forall A\in\mathscr{B}(\Xi).$$

Then

$$\int_\Xi g_{km}(\xi_k)\boldsymbol{\mu}^n(d\xi) = \int_{\Xi_k} g_{km}(t)F_k(dt),\ \ \forall 1\le k\le K, 1\le m\le n,$$

due to (4.5). Since the collection of probability measures $\{\boldsymbol{\mu}^n\}_n$ is tight, by Prokhorov's theorem, there is a convergent subsequence, whose limit is denoted by $\boldsymbol{\mu}^*$. It follows that

$$\int_\Xi g_{km}(\xi_k)\boldsymbol{\mu}^*(d\xi) = \int_{\Xi_k} g_{km}(t)F_k(dt),\ \ \forall 1\le k\le K, m\ge 1,$$

that is, $\boldsymbol{\mu}^*$ has marginals $\{F_k\}_k$. Hence $\int_\Xi f_k(\xi_k)\boldsymbol{\mu}^*(d\xi) = \int_{\Xi_k} f_k(t)F_k(dt)$ for all $f_k \in$

$B(\Xi_k)$. In addition, due to (4.5), we have that $\boldsymbol{\mu}^*$ is primal feasible, and

$$\begin{aligned}
&\int_{\Xi} \ell(\xi)\boldsymbol{\mu}^*(d\xi) \\
=&\int_{\Xi} \big[\ell(\xi) - \sum_{k=1}^{K} f_k^*(\xi_k) - \lambda^* \mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\mu}^*(d\xi) \\
&\qquad\qquad + \int_{\Xi} \big[\sum_{k=1}^{K} f_k^*(\xi_k) + \lambda^* \mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\mu}^*(d\xi) \\
=&\int_{\Xi} \sup_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^{K} f_k^*(\xi_k) - \lambda^* \mathsf{d}_F^p(\xi,\zeta)\Big]\boldsymbol{\nu}(d\zeta) + \lambda^*\theta^p + \sum_{k=1}^{K}\int_{\Xi_k} f_k^*(t)F_k(dt) \\
\geq& v_D.
\end{aligned}$$

□

4.4.1 Data-driven Problem and Size Reduction

Corollary 4.1. *Suppose* $\ell(\xi) = \max_{1\leq m\leq M} a^{m\top}\xi + b^m$ *for some* $a^m \in \mathbb{R}^K$ *and* $b^m \in \mathbb{R}$, *and* $\boldsymbol{\nu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_{\widehat{\xi}^i}$. *Let* $\Xi_k := \{\widehat{\xi}_k^i : \ i = 1,\ldots,n\}$. *Then the dual problem of* (4.2) *is given by*

$$\begin{aligned}
\inf_{\substack{\lambda\geq 0, f_k^i\in\mathbb{R}\\ y_i\in\mathbb{R}}} \Bigg\{ &\lambda\theta^p + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} f_k^i + \frac{1}{n}\sum_{i=1}^{n} y^i : \\
&y^i \geq a^{m\top}(\widehat{\xi}_1^{j_1},\ldots,\widehat{\xi}_K^{j_K}) + b^m - \sum_{k=1}^{K} f_k^{j_k} - \lambda \mathsf{d}_F^p\big((\widehat{\xi}_1^{j_1},\ldots,\widehat{\xi}_K^{j_K}),\widehat{\xi}^i\big), \\
&\forall 1\leq i\leq n,\ \forall 1\leq j_k\leq n,\ \forall 1\leq k\leq K,\ \forall 1\leq m\leq M\Bigg\}.
\end{aligned} \tag{4.6}$$

If, in addition, there exists $\{\mathsf{d}_{F,k}\}_k$ *such that*

$$\mathsf{d}_F^p\big((\widehat{\xi}_1^{j_1},\ldots,\widehat{\xi}_K^{j_K}),\widehat{\xi}^i\big) = \sum_{k=1}^{K} \mathsf{d}_{F,k}(\widehat{\xi}_k^{j_K},\widehat{\xi}_k^i), \ \ \forall i,j_k,\ \forall k,$$

then the above program is equivalent to

$$\inf_{\substack{\lambda\geq 0, f_k^i\in\mathbb{R}\\ y^i, z_k^{im}\in\mathbb{R}}} \Bigg\{\lambda\theta^p + \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n} f_k^i + \frac{1}{n}\sum_{i=1}^{n} y^i :\ y^i \geq b^m + \sum_{k=1}^{K} z_k^{im},\ \forall i, m, \\ z_k^{im} \geq a_k^{m\top}\widehat{\xi}_k^j - f_k^j - \lambda \mathsf{d}_{F,k}(\widehat{\xi}_k^j, \widehat{\xi}_k^i),\ \forall i, m, j, k\Bigg\}. \tag{4.7}$$

Proof of Corollary 4.1. Formulation (4.6) follows directly from Theorem 4.2. Formulation (4.7) follows from the fact that for any additively separable function $g(\widehat{\xi}_1^{j_1}, \ldots, \widehat{\xi}_K^{j_K}) = \sum_{k=1}^{K} g_k(\widehat{\xi}_k^{j_k})$,

$$\max_{j_1,\cdots,j_K} g(\widehat{\xi}_1^{j_1}, \ldots, \widehat{\xi}_K^{j_K}) = \sum_{k=1}^{K} \max_{j} g_k(\widehat{\xi}_k^{j}).$$

□

We remark that (4.7) is of computational importance, as it indicates that when the metric d_F^p is additively separable, by introducing auxiliary variables z_k^{im}, the original problem of size exponential in K admits a reformulation of size linearly growing in dimension K.

4.5 Applications

In this section, we discuss two applications.

4.5.1 Mean-CVaR portfolio selection

We consider a distributionally robust portfolio optimization problem

$$\min_{\beta\in\mathcal{D}} \max_{\boldsymbol{\mu}\in\mathfrak{M}}\ \mathbb{E}_{\boldsymbol{\mu}}[-\beta^\top \boldsymbol{\xi}] + c\cdot \mathrm{CVaR}^{\alpha}_{\boldsymbol{\mu}}[-\beta^\top \boldsymbol{\xi}], \tag{4.8}$$

where $c > 0$, $\mathcal{D} := \left\{\beta \in \mathbb{R}_+^K :\ \sum_{k=1}^{K}\beta_k = 1\right\}$ encodes the vectors of weights of K assets without short-selling, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)^\top$ is the vector of excessive returns over the risk-free rate, and CVaR is the conditional value-at-risk [136] under distribution $\boldsymbol{\mu}$. We use the Fama-French three-factor model [137] to model the asset return. The Fama-French

three-factor model assumes that the excess return of the k-th asset follows the following three-factor model:

$$\boldsymbol{\xi}_k = b_{k1}\boldsymbol{f}_1 + b_{k2}\boldsymbol{f}_2 + b_{k3}\boldsymbol{f}_3 + \boldsymbol{\epsilon}_k, \quad k = 1, \ldots, K,$$

where, the factor $\boldsymbol{f}_1$ are respectively the excess return of the proxy of the market portfolio, which equals the value-weighted return on all NYSE, AMEX and NASDAQ stocks minus the one-month Treasury bill rate; factors $\boldsymbol{f}_2, \boldsymbol{f}_3$ are related to the market capitalizations and and book-to market ratios, more specifically, $\boldsymbol{f}_2$ equals the average return on three small portfolios minus the average return on three big portfolios, and $\boldsymbol{f}_3$ equals the average return on two value portfolios minus the average return on two growth portfolios; b_{k1}, b_{k2}, b_{k3} are the factor loadings of the k-th stock; and ϵ_k is the idiosyncratic noise independent of the three factors, and independent across the stocks.

The parameters are estimated using the three-year daily data of 30 Industry Portfolios from May 1, 2002 to Aug 29, 2005 [138]. We borrow the calibration results from [139] (see Table 4.2), where the factor loadings (b_{k1}, b_{k2}, b_{k3}), $k = 1, \ldots, K$ are i.i.d. drawn from $Normal(\mu_b, \Sigma_b)$, and once generated, they are fixed as constants throughout simulations. The n-period returns of the three factors $(\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3)$ are generated from $Normal(\mu_f, \Sigma_f)$, and the noises are generated from $Gamma(3.3586, 0.1876)$ conditioned on the noise level of at least 0.1950.

Table 4.2: Parameters in the three-factor model

μ_b		Σ_b		μ_f		Σ_f	
0.78282	0.029145	0.023873	0.010184	0.023558	1.2507	-0.034999	-0.20419
0.51803	0.0232873	0.053951	-0.006967	0.012989	-0.034999	0.31564	-0.0022526
0.41003	0.010184	-0.006967	0.086856	0.020714	-0.20419	-0.0022526	0.19303

Note that the objective function of (4.8) can be equivalently written as

$$\min_{\beta \in \mathcal{D}, \tau \in \mathbb{R}} \sup_{\boldsymbol{\mu} \in \mathfrak{M}} \left\{ \mathbb{E}_{\boldsymbol{\mu}} \left[\max_{1 \le m \le M} a_m x^\top \boldsymbol{\xi} + b_m \tau \right] \right\},$$

where $M = 2$, $a_1 = -1$, $a_2 = -1 - c/\alpha$, $b_1 = c$ and $b_2 = c(1 - 1/\alpha)$. We choose $\boldsymbol{\mathcal{C}}_0$ to be the empirical copula in defining $\mathfrak{M}$. In all numerical experiments, we set $\alpha = 0.2$, $c = 10$, $\mathsf{d}(u, v) = ||u - v||_1$. We fix $n = 50$, and vary $K = 10, 50, 100$, corresponding to three regimes $n > K$, $n = K$, and $n < K$. We run the simulation with 200 repetitions. The Wasserstein radius θ is chosen using hold-out cross validation. More specifically, in each repetition, we generate n-period returns, and the n samples are randomly partitioned into a training dataset with 70% data and a validation set with 30% data. We solve problem (4.8) using the training dataset for different choices of θ, and choose the one that has the best out-of-sample performance using validation dataset. Then we resolve problem (4.8) using the all n samples, and the out-of-sample performance of the optimal solution is evaluated using an independent testing dataset with 10^6 samples.

We compare our approach with two other approaches, sample average approximation (SAA) method, and DRSO with $\mathcal{W}_1$-Wasserstein ball considered in [24], in which there is no constraints on the marginal distributions and the ball is centered at the empirical distribution instead of the copula. Note that our numerical setting is similar to the one in [24], expect that we generate random asset returns based on the three-factor model whose parameters are calibrated using real data. The box plot of the results is shown in Figure 4.3.

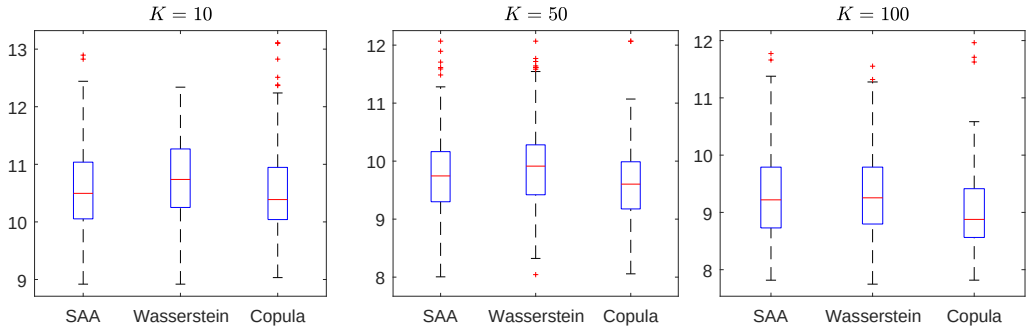


Figure 4.3: Out-of-sample performances of three approaches

We observe that the DRSO with Wasserstein ball does not have a superior performance over SAA method, and is actually even worse in relatively low dimensional setting when

$K \leq n$. Possible explanation of this is that variations of the uncertain asset returns are not that big, so SAA already has a relatively good performance especially in low-dimensional setting, whereas DRSO with Wasserstein ball only provide a conservative solution. Nevertheless, our proposed Copula approach seems to perform better when the dimensional K becomes larger. Note that in our experiments, samples of size $n = 50$ already provide a rather accurate estimate of the one-dimensional marginal distribution. By constraining the marginal distributions and building a ball around the empirical copula, our approach obtain a more robust (comparing to SAA) yet less conservative solution (comparing to Wasserstein ball), and this effect becomes more apparent in high dimensions.

4.5.2 Nonparametric density estimation with extra marginal data

We focus on the copula density estimation in the second step above, and we are interested in nonparametric estimation. The following setup is based on [140]. The domain $[0,1]^2$ is partitioned into $M \times M$ rectangle cells with equal size. For each cell $(u_1^{k_1}, u_2^{k_2})$, $k_1, k_2 = 1, \ldots, M$, denote by $\boldsymbol{C}^0_{k_1,k_2}$ the empirical relative frequency of observations $\{(\hat{u}_1^i, \hat{u}_2^i)\}_{i=1}^n$ falling in this cell, and define x_{k_1,k_2} to be the probability mass of this cell that we are going to estimate. Then the maximum likelihood estimation is given by

$$\min_{\beta \in \mathcal{D}} \ \mathbb{E}_{\boldsymbol{C}^0}[-\log(\beta(\boldsymbol{u}))], \tag{4.9}$$

where

$$\mathcal{D} := \left\{ \beta \in \mathbb{R}_+^{M \times M} : \ \sum_{k_1} \beta_{k_1,k_2} = \sum_{k_2} \beta_{k_1,k_2} = \frac{1}{M} \right\}.$$

In [140], it is proposed to consider a total variation penalized likelihood

$$\min_{\beta \in \mathcal{D}} \ \mathbb{E}_{\boldsymbol{C}^0}[-\log(\beta(\boldsymbol{u}))] + \lambda \sum_{k_1,k_2=1}^{M} \sqrt{(\beta_{k_1+1,k_2} - \beta_{k_1,k_2})^2 + (\beta_{k_1,k_2+1} - \beta_{k_1,k_2})^2}.$$

Here we propose another approach based on our distributionally robust framework.

Consider

$$\min_{\beta\in\mathcal{D}} \max_{\boldsymbol{C}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{C}}\big[-\log(\beta(\boldsymbol{u}))\big], \tag{4.10}$$

where $\mathfrak{M}$ is a ball of subcopulas centered at $\boldsymbol{C}_0$. Using our duality result, the problem above can be reformulated as a convex programming

$$\min_{\substack{\beta\in\mathcal{D},\lambda\geq 0\\ f_1^{k_1},f_2^{k_2},y}} \Bigg\{\lambda\theta + \frac{1}{M}\sum_{k_1=1}^{M} f_1^{k_1} + \frac{1}{M}\sum_{k_2=1}^{M} f_2^{k_2} + \frac{1}{n}\sum_{i=1}^{n} y_i + \sum_{k_1,k_2=1}^{M} x_{k_1,k_2}^2 :$$
$$y_i \geq -\log(\beta_{k_1,k_2}) - f_1^{k_1} - f_2^{k_2} - \lambda\cdot||(u_1^{k_1},u_2^{k_2}) - (\hat{u}_1^i,\hat{u}_2^i)||_1,\ \forall i,k_1,k_2\Bigg\}.$$

In our experiment, we use a dataset in Example 4.6. We compare our approach with total variation penalized likelihood estimation proposed in [140], which is, to the best of our knowledge, the only method that fores the marginal constraints on the copula (Many other kernel/wavelets-based approach actually do not provide an estimator that satisfies the marginal requirement for a copula). In our experiment, we set $M = 32$. Since the real dataset is very small, we here only provide a qualitative comparison for the copula density estimators.

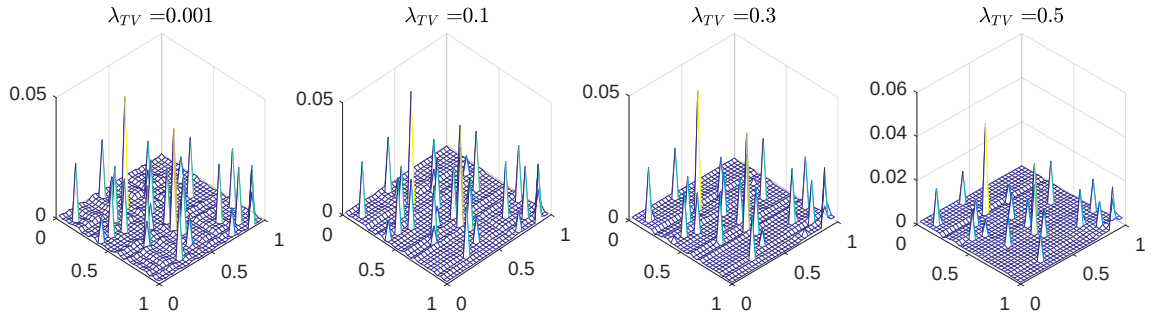


Figure 4.4: Copula density estimator using TV penalized maximum likelihood

Figure 4.4 and 4.5 show the estimators yielding from the two approaches with different tuning parameters. It is obvious that they differ a lot. In particular, the density estimator using total variation penalized likelihood estimation proposed in [140] has disconnected support, which seems unrealistic. In contrast, our density estimator is smoother and seems

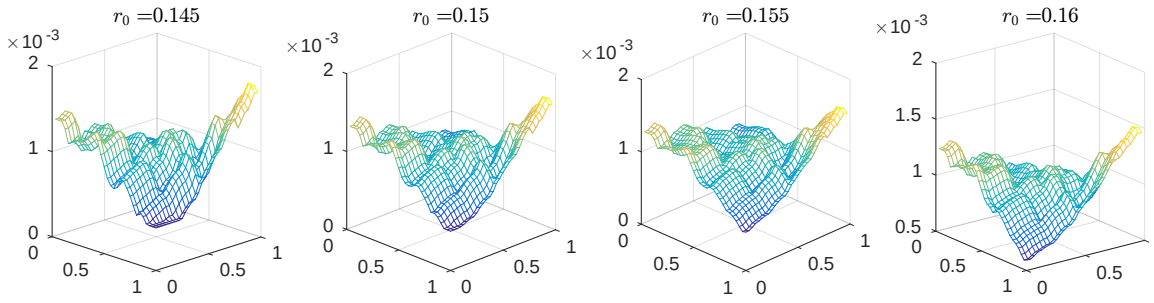


Figure 4.5: Copula density estimator using Wasserstein-based distributionally robust method

to be more reasonable using only a small dataset.

4.6 Concluding remarks

In this chapter, we proposed a distributionally robust framework for decision-making under uncertainty when the marginal distributions are fixed. We chose Wasserstein distance to measure the closeness between the considered dependence structure and some nominal model. We used several illustrative examples to show its advantages over previous work on divergence-based approach. Our computational examples on portfolio selection and density estimation show that, for high-dimensional data-driven problems, namely, problems in which the sample size is much less than the number of unknown parameters, our approach outperforms the conventional approaches.

Appendices

APPENDIX A

APPENDIX FOR CHAPTER 2

A.1 Auxiliary results

Lemma A.1. *Consider any $p \geq 1$ and any $\varepsilon > 0$. Then there exists $C_p(\varepsilon) \geq 1$ such that*

$$(x+y)^p \;\; \leq \;\; (1+\varepsilon)x^p + C_p(\varepsilon)y^p$$

for all $x, y \geq 0$.

Lemma A.2. *Consider any $\zeta^0 \in \Xi$. Then for any $\lambda > \lambda_1 > \kappa$, there exists a constant $C > 0$ such that*

$$\frac{\lambda - \lambda_1}{2}\overline{D}(\lambda, \zeta) \;\; \leq \;\; \Phi(\lambda, \zeta) - \Phi(\lambda_1, \zeta^0) + \lambda_1 C \mathsf{d}^p(\zeta, \zeta^0)$$

for all $\zeta \in \Xi$.

Lemma A.3. *Suppose (A.20) holds and the constant L, M are defined in (A.21). Then for the vector field defined in (A.22), it holds that $||F(z) - F(z')||_{Z,*} \leq L||z - z'||_Z + M$ for all $z, z' \in Z$.*

Lemma A.4. *Let C be a Borel set in Ξ with nonempty boundary ∂C. Then for any $\varepsilon > 0$, there exists a Borel map $T_\varepsilon : \partial C \to \Xi \setminus \mathrm{cl}(C)$ such that $\mathsf{d}(\xi, T_\varepsilon(\xi)) < \varepsilon$ for all $\xi \in \partial C$.*

A.2 Proofs

A.2.1 Proofs of Lemmas

Proof of Lemma 2.1. Let (u_0, v_0) be any feasible solution for the maximization problem in (2.2). For any $t \in \mathbb{R}$ and any $\xi, \zeta \in \Xi$, let $u_t(\xi) := u_0(\xi) + t$ and $v_t(\zeta) := v_0(\zeta) - t$.

Then it follows that $u_t(\xi) + v_t(\zeta) \le \mathsf{d}^p(\xi, \zeta)$ for all $\xi, \zeta \in \Xi$, and

$$\int_\Xi u_t(\xi)\boldsymbol{\mu}(d\xi) + \int_\Xi v_t(\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \int_\Xi u_0(\xi)\boldsymbol{\mu}(d\xi) + \int_\Xi v_0(\zeta)\boldsymbol{\nu}(d\zeta) + t[\boldsymbol{\mu}(\Xi) - \boldsymbol{\nu}(\Xi)].$$

Since $\boldsymbol{\mu}(\Xi) \neq \boldsymbol{\nu}(\Xi)$,

$$\sup_{t\in\mathbb{R}} \left\{ \int_\Xi u_t(\xi)\boldsymbol{\mu}(d\xi) + \int_\Xi v_t(\zeta)\boldsymbol{\nu}(d\zeta) \right\} \;=\; \infty,$$

and thus $\mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \infty$. □

Proof of Lemma 2.2. (i) We prove the result by contradiction. Suppose that for some $\zeta^0, \zeta^1 \in \Xi$, it holds that

$$\kappa^0 \;:=\; \limsup_{\xi\in\Xi\,:\,d(\xi,\zeta^0)\to\infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta^0)\}}{\mathsf{d}^p(\xi, \zeta^0)} \;<\; \kappa^1 \;:=\; \limsup_{\xi\in\Xi\,:\,d(\xi,\zeta^1)\to\infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta^1)\}}{\mathsf{d}^p(\xi, \zeta^1)}$$

($\kappa^1 = \infty$ is allowed). Choose any $\varepsilon \in (0, \kappa^1 - \kappa^0)$. Then there exists an R such that for all ξ with $d(\xi, \zeta^0) > R$ it holds that

$$\begin{aligned}
\ell(\xi) - \ell(\zeta^1) &= \ell(\xi) - \ell(\zeta^0) + \ell(\zeta^0) - \ell(\zeta^1) \\
&\le \max\{0, \ell(\xi) - \ell(\zeta^0)\} + \ell(\zeta^0) - \ell(\zeta^1) \\
&< (\kappa^0 + \varepsilon)\mathsf{d}^p(\xi, \zeta^0) + [\ell(\zeta^0) - \ell(\zeta^1)] \\
&\le (\kappa^0 + \varepsilon)\left[\mathsf{d}(\xi, \zeta^1) + d(\zeta^1, \zeta^0)\right]^p + [\ell(\zeta^0) - \ell(\zeta^1)]
\end{aligned}$$

Since $\kappa^1 > 0$, it follows that

$$\begin{aligned}
\kappa^1 &= \limsup_{\xi\in\Xi\,:\,d(\xi,\zeta^1)\to\infty} \frac{\ell(\xi) - \ell(\zeta^1)}{\mathsf{d}^p(\xi, \zeta^1)} \\
&\le \limsup_{\xi\in\Xi\,:\,d(\xi,\zeta^1)\to\infty} \frac{(\kappa^0 + \varepsilon)\left[\mathsf{d}(\xi, \zeta^1) + d(\zeta^1, \zeta^0)\right]^p + [\ell(\zeta^0) - \ell(\zeta^1)]}{\mathsf{d}^p(\xi, \zeta^1)} \\
&= \kappa^0 + \varepsilon \;<\; \kappa^1,
\end{aligned}$$

which is a contradiction.

(ii) First we show that if there exists $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$, then $\kappa < \infty$. Let $\kappa^0 := 0$ if Ξ is bounded, and let

$$\kappa^0 \;:=\; \limsup_{\xi \in \Xi\,:\,\mathsf{d}^p(\xi,\zeta^0)\to\infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta^0)\}}{\mathsf{d}^p(\xi, \zeta^0)} \;\leq\; L \;<\; \infty$$

if Ξ is unbounded. If Ξ is unbounded, then it follows from (i) that

$$\kappa^0 \;=\; \limsup_{\xi \in \Xi\,:\,\mathsf{d}^p(\xi,\zeta)\to\infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta)\}}{\mathsf{d}^p(\xi, \zeta)} \quad \forall\, \zeta \in \Xi. \tag{A.1}$$

We are going to show that $\int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) > -\infty$ for all $\lambda > \kappa^0$, and therefore $\kappa \leq \kappa^0 < \infty$.

First we show that $\Phi(\lambda, \zeta) > -\infty$ for any $\lambda > \kappa^0$ and $\zeta \in \Xi$. If Ξ is bounded, then choose any $R(\zeta) > 0$ such that $\mathsf{d}^p(\xi, \zeta) \leq R(\zeta)$ for all $\xi \in \Xi$. If Ξ is unbounded, then it follows from (A.1) that for any $\zeta \in \Xi$, there is a $R(\zeta) > 0$ such that for all $\xi \in \Xi$ with $\mathsf{d}^p(\xi, \zeta) > R(\zeta)$, it holds that

$$\frac{\ell(\xi) - \ell(\zeta)}{\mathsf{d}^p(\xi, \zeta)} \;<\; \frac{\lambda + \kappa^0}{2},$$

that is, $(\frac{\lambda+\kappa^0}{2})\mathsf{d}^p(\xi, \zeta) - \ell(\xi) > -\ell(\zeta)$. Thus, for all $\xi \in \Xi$ with $\mathsf{d}^p(\xi, \zeta) > R(\zeta)$, it holds that

$$\begin{aligned}
\lambda\mathsf{d}^p(\xi, \zeta) - \ell(\xi) \;&=\; \frac{\lambda + \kappa^0}{2}\mathsf{d}^p(\xi, \zeta) - \ell(\xi) + \frac{\lambda - \kappa^0}{2}\mathsf{d}^p(\xi, \zeta) \\
&>\; -\ell(\zeta) + \frac{\lambda - \kappa^0}{2}R(\zeta),
\end{aligned}$$

and hence

$$\inf_{\xi\in\Xi}\left\{\lambda\mathsf{d}^p(\xi, \zeta) - \ell(\xi) \,:\, \mathsf{d}^p(\xi, \zeta) > R(\zeta)\right\} \;\geq\; -\ell(\zeta) + \frac{\lambda - \kappa^0}{2}R(\zeta) \;>\; -\infty.$$

Also, by assumption, for any $\xi \in \Xi$ it holds that

$$\begin{aligned}\ell(\xi) - \ell(\zeta^0) &\leq L\mathsf{d}^p(\xi, \zeta^0) + M \\ &\leq L[\mathsf{d}(\xi, \zeta) + \mathsf{d}(\zeta, \zeta^0)]^p + M \\ &\leq 2^{p-1}L[\mathsf{d}^p(\xi, \zeta) + \mathsf{d}^p(\zeta, \zeta^0)] + M\end{aligned}$$

where the second inequality follows from the elementary inequality $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ for any $a, b \geq 0$ and $p \geq 1$. Thus

$$\begin{aligned}\inf_{\xi\in\Xi}\left\{\lambda\mathsf{d}^p(\xi, \zeta) - \ell(\xi) \ : \ \mathsf{d}^p(\xi, \zeta) \leq R(\zeta)\right\} &\geq \inf_{\xi\in\Xi}\left\{ - \ell(\xi) \ : \ \mathsf{d}^p(\xi, \zeta) \leq R(\zeta)\right\} \\ &\geq -\ell(\zeta^0) - 2^{p-1}LR(\zeta) - 2^{p-1}L\mathsf{d}^p(\zeta, \zeta^0) - M \\ &> -\infty.\end{aligned}$$

Therefore, $\Phi(\lambda, \zeta) > -\infty$ for all $\zeta \in \Xi$ and $\lambda > \kappa^0$.

Next we show that $\int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) > -\infty$ for any $\lambda > \kappa^0$. Consider any $\lambda_1 \in (\kappa^0, \lambda)$ and any $\zeta^0 \in \Xi$. It follows from Lemma A.2 that there is a constant C such that

$$\Phi(\lambda, \zeta) \ \geq \ \frac{\lambda - \lambda_1}{2}\overline{D}(\lambda, \zeta) + \Phi(\lambda_1, \zeta^0) - C\mathsf{d}^p(\zeta, \zeta^0) \ \geq \ \Phi(\lambda_1, \zeta^0) - C\mathsf{d}^p(\zeta, \zeta^0).$$

Thus

$$\int_\Xi \Phi(\lambda, \zeta)\boldsymbol{\nu}(d\zeta) \ \geq \ \Phi(\lambda_1, \zeta^0) - C\int_\Xi \mathsf{d}^p(\zeta, \zeta^0)\boldsymbol{\nu}(d\zeta) \ > \ -\infty.$$

Therefore $\kappa \leq \kappa^0 < \infty$.

Next we show that if there does not exist $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$, then $\kappa = \infty$. First, observe that if there exists $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$, then for any $\xi, \zeta \in \Xi$ it

holds that

$$\begin{aligned}\ell(\xi) - \ell(\zeta) &= \ell(\xi) - \ell(\zeta^0) + \ell(\zeta^0) - \ell(\zeta) \\ &\leq L\mathsf{d}^p(\xi, \zeta^0) + M + \ell(\zeta^0) - \ell(\zeta) \\ &\leq L[\mathsf{d}(\xi, \zeta) + d(\zeta, \zeta^0)]^p + M + \ell(\zeta^0) - \ell(\zeta) \\ &\leq 2^{p-1}L[\mathsf{d}^p(\xi, \zeta) + \mathsf{d}^p(\zeta, \zeta^0)] + M + \ell(\zeta^0) - \ell(\zeta)\end{aligned}$$

It follows that there exists $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$ if and only if there exists $L' := 2^{p-1}L \geq 0$ and $M(\zeta) := 2^{p-1}L\mathsf{d}^p(\zeta, \zeta^0) + M + \ell(\zeta^0) - \ell(\zeta) \in L^1(\boldsymbol{\nu})$ such that $\ell(\xi) - \ell(\zeta) \leq L'\mathsf{d}^p(\xi, \zeta) + M(\zeta)$ for all $\xi, \zeta \in \Xi$, that is, there exists $L' \geq 0$ and $M(\zeta) \in L^1(\boldsymbol{\nu})$ such that $-\ell(\zeta) - M(\zeta) \leq \inf_{\xi \in \Xi} \{L'\mathsf{d}^p(\xi, \zeta) - \ell(\xi)\}$ for all $\zeta \in \Xi$. Therefore, if there does not exist $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$, then for any $\lambda \geq 0$ it holds that

$$\inf_{\xi \in \Xi} \{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\} \quad \notin \quad L^1(\boldsymbol{\nu})$$

which implies that $\kappa = \infty$.

(iii) It was established in the proof of (ii) that if $\kappa < \infty$ then there exists $\zeta^0 \in \Xi$ and $L, M > 0$ such that $\ell(\xi) - \ell(\zeta^0) \leq L\mathsf{d}^p(\xi, \zeta^0) + M$ for all $\xi \in \Xi$, and then

$$\kappa \;\leq\; \kappa^0 \;:=\; \limsup_{\xi \in \Xi \,:\, d(\xi, \zeta^0) \to \infty} \frac{\max\{0, \ell(\xi) - \ell(\zeta^0)\}}{\mathsf{d}^p(\xi, \zeta^0)}$$

Next we show that $\kappa \geq \kappa^0$. If $\kappa^0 = 0$, then it follows from the definition of κ that $\kappa \geq \kappa^0$. Next, suppose that $\kappa^0 > 0$, and consider any $\lambda \in [0, \kappa^0)$. We will show that $\inf_{\xi \in \Xi} \{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\} = -\infty$ for all $\zeta \in \Xi$. If $\lambda = 0$, then it follows from $\kappa^0 > 0$ that $\inf_{\xi \in \Xi} \{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\} = -\infty$ for all ζ. Next, consider any $\lambda \in (0, \kappa^0)$, any $\zeta \in \Xi$, any $M > 0$, any $\lambda_2 \in (\lambda, \kappa^0)$, and any $\varepsilon \in (0, (\lambda_2 - \lambda)/\lambda)$. Since $[\mathsf{d}(\xi, \zeta^0) + \mathsf{d}(\zeta^0, \zeta)]^p / \mathsf{d}^p(\xi, \zeta^0) \to 1$ as $\mathsf{d}^p(\xi, \zeta^0) \to \infty$, it follows that there exists $R_1 > 0$ such

that $\mathsf{d}^p(\xi,\zeta)/\mathsf{d}^p(\xi,\zeta^0) \leq [\mathsf{d}(\xi,\zeta^0)+\mathsf{d}(\zeta^0,\zeta)]^p/\mathsf{d}^p(\xi,\zeta^0) \leq 1+\varepsilon$ for all $\xi \in \Xi$ such that $\mathsf{d}^p(\xi,\zeta^0) > R_1$. Choose any $R > \max\{R_1, [M-\ell(\zeta^0)]/(\lambda_2-\lambda-\lambda\varepsilon)\}$. It follows from the definition of κ^0 that there exists $\xi \in \Xi$ such that $\mathsf{d}^p(\xi,\zeta^0) > R$ and

$$\begin{aligned} \ell(\xi)-\ell(\zeta^0) &> \lambda_2 \mathsf{d}^p(\xi,\zeta^0) \\ &= \lambda \mathsf{d}^p(\xi,\zeta^0) + (\lambda_2-\lambda)\mathsf{d}^p(\xi,\zeta^0) \\ &\geq \lambda \mathsf{d}^p(\xi,\zeta) + (\lambda_2-\lambda-\lambda\varepsilon)\mathsf{d}^p(\xi,\zeta^0) \\ \Rightarrow \quad \lambda \mathsf{d}^p(\xi,\zeta)-\ell(\xi) &< -\ell(\zeta^0)-(\lambda_2-\lambda-\lambda\varepsilon)R < -M \end{aligned}$$

Thus, $\inf_{\xi\in\Xi}\{\lambda \mathsf{d}^p(\xi,\zeta)-\ell(\xi)\} = -\infty$ for all ζ, and hence $\int_\Xi \inf_{\xi\in\Xi}\{\lambda \mathsf{d}^p(\xi,\zeta)-\ell(\xi)\}\,\boldsymbol{\nu}(d\zeta) = -\infty$ for all $\lambda \in [0,\kappa^0)$. Therefore, $\kappa \geq \kappa^0$. Next, recall that (i) established that κ^0 does not depend on the choice of ζ^0, and therefore the result follows. □

Proof of Lemma 2.3.

(i) For any $\zeta \in \Xi$, $\Phi(\cdot,\zeta)$ is the infimum of nondecreasing functions. Thus $\Phi(\cdot,\zeta)$ is nondecreasing for all $\zeta \in \Xi$. Also, for any $\zeta \in \Xi$, $\Phi(\cdot,\zeta)$ is the infimum of continuous functions. Thus $\Phi(\cdot,\zeta)$ is upper-semicontinuous for all $\zeta \in \Xi$. Consider any sequence $\{\lambda_n\}_n$ such that $\lambda_n \downarrow \kappa$ as $n \to \infty$. Since $\int_\Xi \Phi(\lambda_n,\zeta)\boldsymbol{\nu}(d\zeta) > -\infty$, it holds that there is a set $B_n \in \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ such that $\boldsymbol{\nu}(B_n) = 1$ and $\Phi(\lambda_n,\zeta) > -\infty$ for all $\zeta \in B_n$. Then it follows from $\Phi(\cdot,\zeta)$ being nondecreasing that $\Phi(\lambda,\zeta) > -\infty$ for all $\lambda \geq \lambda_n$ and all $\zeta \in B_n$. Let $B := \cap_n B_n$. Then $B \in \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$, and $\boldsymbol{\nu}(B) = 1$, and $\Phi(\lambda,\zeta) > -\infty$ for all $\lambda > \kappa$ and all $\zeta \in B$. Since $\Phi(\cdot,\zeta)$ is the infimum of affine functions of λ, and $\Phi(\lambda,\zeta) < \infty$ for all $\lambda \geq 0$ and all $\zeta \in \Xi$, and $\Phi(\lambda,\zeta) > -\infty$ for all $\lambda > \kappa$ and all $\zeta \in B$, it follows that $\Phi(\cdot,\zeta)$ is concave on $[0,\infty)$ for all $\zeta \in B$.

For the second part, consider any $\lambda_2 > \lambda_1$ and any $\zeta \in \Xi$ such that $\Phi(\lambda_i,\zeta) > -\infty$ for $i = 1,2$. For any $\delta_i > 0$, consider any $\xi_i^{\delta_i} \in \Xi$ such that $\lambda_i \mathsf{d}^p(\xi_i^{\delta_i},\zeta) - \ell(\xi_i^{\delta_i}) \leq \Phi(\lambda_i,\zeta)+\delta_i$

for $i = 1, 2$. It follows that

$$\begin{aligned}
\lambda_2 \mathsf{d}^p(\xi_2^{\delta_2}, \zeta) - \ell(\xi_2^{\delta_2}) &\leq \Phi(\lambda_2, \zeta) + \delta_2 \\
&\leq \lambda_2 \mathsf{d}^p(\xi_1^{\delta_1}, \zeta) - \ell(\xi_1^{\delta_1}) + \delta_2 \\
&= (\lambda_2 - \lambda_1)\mathsf{d}^p(\xi_1^{\delta_1}, \zeta) + \lambda_1 \mathsf{d}^p(\xi_1^{\delta_1}, \zeta) - \ell(\xi_1^{\delta_1}) + \delta_2 \\
&\leq (\lambda_2 - \lambda_1)\mathsf{d}^p(\xi_1^{\delta_1}, \zeta) + \Phi(\lambda_1, \zeta) + \delta_1 + \delta_2 \\
&\leq (\lambda_2 - \lambda_1)\mathsf{d}^p(\xi_1^{\delta_1}, \zeta) + \lambda_1 \mathsf{d}^p(\xi_2^{\delta_2}, \zeta) - \ell(\xi_2^{\delta_2}) + \delta_1 + \delta_2 \\
\Rightarrow \quad \mathsf{d}^p(\xi_2^{\delta_2}, \zeta) - \frac{\delta_2}{\lambda_2 - \lambda_1} &\leq \mathsf{d}^p(\xi_1^{\delta_1}, \zeta) + \frac{\delta_1}{\lambda_2 - \lambda_1}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad & \sup_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) - \frac{\delta_2}{\lambda_2 - \lambda_1} \ : \ \lambda_2 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda_2, \zeta) + \delta_2 \right\} \\
& \leq \inf_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) + \frac{\delta_1}{\lambda_2 - \lambda_1} \ : \ \lambda_1 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda_1, \zeta) + \delta_1 \right\} \\
\Rightarrow \quad & \limsup_{\delta_2 \downarrow 0} \left\{ \sup_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) - \frac{\delta_2}{\lambda_2 - \lambda_1} \ : \ \lambda_2 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda_2, \zeta) + \delta_2 \right\} \right\} \\
& \leq \liminf_{\delta_1 \downarrow 0} \left\{ \inf_{\xi \in \Xi} \left\{ \mathsf{d}^p(\xi, \zeta) + \frac{\delta_1}{\lambda_2 - \lambda_1} \ : \ \lambda_1 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda_1, \zeta) + \delta_1 \right\} \right\} \\
\Rightarrow \quad & \overline{D}(\lambda_2, \zeta) \ \leq \ \underline{D}(\lambda_1, \zeta).
\end{aligned}$$

Also, it follows from the definition of $\overline{D}$ and $\underline{D}$ that $\underline{D}(\lambda_1, \zeta) \leq \overline{D}(\lambda_1, \zeta)$.

(ii) It follows from the definition of Φ that for all $\xi, \zeta \in \Xi$ it holds that

$$\ell(\xi) \ \leq \ \lambda_1 \mathsf{d}^p(\xi, \zeta) - \Phi(\lambda_1, \zeta).$$

Also, for every $\xi \in \Xi$ that satisfies $\lambda_2 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \leq \Phi(\lambda_2, \zeta) + \delta$ for some $\delta \geq 0$, it holds that

$$\lambda_2 \mathsf{d}^p(\xi, \zeta) - \ell(\xi) - \delta \ \leq \ -\ell(\zeta).$$

Combining the two inequalities above yields that

$$\begin{aligned}
& \lambda_2 \mathsf{d}^p(\xi,\zeta) + \ell(\zeta) - \delta \;\leq\; \lambda_1 \mathsf{d}^p(\xi,\zeta) - \Phi(\lambda_1,\zeta) \\
\Rightarrow \quad & (\lambda_2-\lambda_1)\mathsf{d}^p(\xi,\zeta) - \delta \;\leq\; -\ell(\zeta) - \Phi(\lambda_1,\zeta) \\
\Rightarrow \quad & \limsup_{\delta\downarrow 0} \left\{ \sup_{\xi\in\Xi} \left\{ (\lambda_2-\lambda_1)\mathsf{d}^p(\xi,\zeta) - \delta \;:\; \lambda_2 \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \leq \Phi(\lambda_2,\zeta) + \delta \right\} \right\} \\
& \quad \leq\; -\ell(\zeta) - \Phi(\lambda_1,\zeta) \\
\Rightarrow \quad & (\lambda_2-\lambda_1)\overline{D}(\lambda_2,\zeta) \;\leq\; -\ell(\zeta) - \Phi(\lambda_1,\zeta)
\end{aligned}$$

(iii) Consider any $\zeta \in B$ and any $\lambda_2 > \lambda_1 > \kappa$. For any $\delta > 0$, choose any $\xi_i^\delta \in \Xi$ such that $\lambda_i \mathsf{d}^p(\xi_i^\delta,\zeta) - \ell(\xi_i^\delta) \leq \Phi(\lambda_i,\zeta) + \delta$ for $i = 1, 2$. Then

$$\Phi(\lambda_1,\zeta) - \Phi(\lambda_2,\zeta) \;\leq\; \lambda_1 \mathsf{d}^p(\xi_2^\delta,\zeta) - \ell(\xi_2^\delta) - \Big[\lambda_2 \mathsf{d}^p(\xi_2^\delta,\zeta) - \ell(\xi_2^\delta)\Big] + \delta \;=\; (\lambda_1-\lambda_2)\mathsf{d}^p(\xi_2^\delta,\zeta) + \delta.$$

Similarly, $\Phi(\lambda_2,\zeta) - \Phi(\lambda_1,\zeta) \leq (\lambda_2-\lambda_1)\mathsf{d}^p(\xi_1^\delta,\zeta) + \delta$. It follows that

$$\mathsf{d}^p(\xi_2^\delta,\zeta) - \frac{\delta}{\lambda_2-\lambda_1} \;\leq\; \frac{\Phi(\lambda_2,\zeta) - \Phi(\lambda_1,\zeta)}{\lambda_2-\lambda_1} \;\leq\; \mathsf{d}^p(\xi_1^\delta,\zeta) + \frac{\delta}{\lambda_2-\lambda_1}.$$

Then it follows from the definitions of $\overline{D}$ and $\underline{D}$ that

$$\overline{D}(\lambda_2,\zeta) \;\leq\; \frac{\Phi(\lambda_2,\zeta) - \Phi(\lambda_1,\zeta)}{\lambda_2-\lambda_1} \;\leq\; \underline{D}(\lambda_1,\zeta).$$

Since $\lambda_1 \in \mathrm{int}(\mathrm{dom}(\Phi(\cdot,\zeta)))$, there is a $\lambda_0 < \lambda_1$ such that $\Phi(\cdot,\zeta)$ is finite-valued and concave on (λ_0,∞), and the left and right derivatives $\partial\Phi(\lambda,\zeta)/\partial_{\lambda\pm}$ exist for all $\lambda \in (\lambda_0,\infty)$. Setting $\lambda_2 = \lambda$ and letting $\lambda_1 \uparrow \lambda$, it follows that

$$\overline{D}(\lambda,\zeta) \;\leq\; \frac{\partial\Phi(\lambda,\zeta)}{\partial\lambda-} \;\leq\; \lim_{\lambda_1\uparrow\lambda} \underline{D}(\lambda_1,\zeta).$$

Similarly, setting $\lambda_1 = \lambda$ and letting $\lambda_2 \downarrow \lambda$ in the inequality above, it follows that

$$\lim_{\lambda_2 \downarrow \lambda} \overline{D}(\lambda_2, \zeta) \;\leq\; \frac{\partial \Phi(\lambda, \zeta)}{\partial \lambda +} \;\leq\; \underline{D}(\lambda, \zeta).$$

□

Proof of Lemma 2.4.

(i) By Definition 1.11 in [72], $\boldsymbol{\nu}$ has an extension, still denoted by $\boldsymbol{\nu}$, such that the measure space $(\Xi, \mathscr{B}_{\boldsymbol{\nu}}, \boldsymbol{\nu})$ is complete. Note that for any $b \in \mathbb{R}$, it holds that

$$\begin{aligned}\{\zeta \in \Xi \,:\, \Phi(\lambda, \zeta) < b\} &= \{\zeta \in \Xi \,:\, \exists\, \xi \in \Xi \text{ such that } \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) < b\} \\ &= \pi^2\big(\{(\xi, \zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) < b\}\big).\end{aligned}$$

Note that the set $\{(\xi, \zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) < b\}$ on the right side is measurable. Since (Ξ, d) is Polish, it follows from the measurable projection theorem (cf. Theorem 8.3.2 in [141]), that $\Phi(\lambda, \cdot)$ is $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-measurable.

Define functions $\overline{C}, \underline{C}$ by

$$\begin{aligned}\overline{C}(\lambda, \zeta, \delta) &:= \sup_{\xi \in \Xi}\{\mathsf{d}^p(\xi, \zeta) \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \,<\, \Phi(\lambda, \zeta) + \delta\} \\ \underline{C}(\lambda, \zeta, \delta) &:= \inf_{\xi \in \Xi}\{\mathsf{d}^p(\xi, \zeta) \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \,<\, \Phi(\lambda, \zeta) + \delta\}.\end{aligned}$$

For any $b \in \mathbb{R}$ it holds that

$$\begin{aligned}&\{\zeta \in \Xi \,:\, \overline{C}(\lambda, \zeta, \delta) > b\} \\ =\; &\{\zeta \in \Xi \,:\, \exists\, \xi \in \Xi \text{ such that } \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \,<\, \Phi(\lambda, \zeta) + \delta,\; \mathsf{d}^p(\xi, \zeta) \,>\, b\} \\ =\; &\pi^2\big(\{(\xi, \zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi) \,<\, \Phi(\lambda, \zeta) + \delta,\; \mathsf{d}^p(\xi, \zeta) \,>\, b\}\big)\end{aligned}$$

and thus it follows from the measurable projection theorem that $\overline{C}(\lambda, \cdot, \delta)$ is $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-

measurable. Similarly,

$$\begin{aligned}
&\{\zeta \in \Xi \,:\, \underline{C}(\lambda,\zeta,\delta) < b\} \\
=\;& \{\zeta \in \Xi \,:\, \exists\, \xi \in \Xi \text{ such that } \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,<\, \Phi(\lambda,\zeta) + \delta,\ \mathsf{d}^p(\xi,\zeta) < b\} \\
=\;& \pi^2\big(\{(\xi,\zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,<\, \Phi(\lambda,\zeta) + \delta,\ \mathsf{d}^p(\xi,\zeta) < b\}\big)
\end{aligned}$$

and thus $\underline{C}(\lambda,\cdot,\delta)$ is $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-measurable.

Next, note that $\overline{D}(\lambda,\cdot) = \limsup_{\delta\downarrow 0} \overline{C}(\lambda,\cdot,\delta)$ and $\underline{D}(\lambda,\cdot) = \liminf_{\delta\downarrow 0} \underline{C}(\lambda,\cdot,\delta)$ are also $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-measurable, because measurability is preserved under $\limsup$ and $\liminf$.

For any $b \in \mathbb{R}$ it holds that

$$\begin{aligned}
&\{\zeta \in \Xi \,:\, \overline{D}_0(\lambda,\zeta) > b\} \\
=\;& \{\zeta \in \Xi \,:\, \exists\, \xi \in \Xi \text{ such that } \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,=\, \Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta) \,>\, b\} \\
=\;& \pi^2\big(\{(\xi,\zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,=\, \Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta) \,>\, b\}\big)
\end{aligned}$$

and thus it follows from the measurable projection theorem that $\overline{D}_0(\lambda,\cdot)$ is $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-measurable. Similarly,

$$\begin{aligned}
&\{\zeta \in \Xi \,:\, \underline{D}_0(\lambda,\zeta) < b\} \\
=\;& \{\zeta \in \Xi \,:\, \exists\, \xi \in \Xi \text{ such that } \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,=\, \Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta) \,<\, b\} \\
=\;& \pi^2\big(\{(\xi,\zeta) \in \Xi \times \Xi \,:\, \lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \,=\, \Phi(\lambda,\zeta),\ \mathsf{d}^p(\xi,\zeta) \,<\, b\}\big)
\end{aligned}$$

and thus $\underline{D}_0(\lambda,\cdot)$ is $(\mathscr{B}_{\boldsymbol{\nu}}, \mathscr{B}(\mathbb{R}))$-measurable.

(ii) For each $\zeta \in \Xi$, it follows from the measurability of ℓ and $\mathsf{d}^p(\cdot,\zeta)$ that $\overline{F}(\lambda,\zeta)$ and $\underline{F}(\lambda,\zeta)$ are in $\mathscr{B}(\Xi)$. Since (Ξ,d) is Polish and $\boldsymbol{\nu}$ is a complete finite measure, it follows from Aumann's measurable selection theorem (see, e.g. Theorem 18.26 in [142]) that $\boldsymbol{\nu}$-measurable selections $\overline{T}(\lambda,\cdot), \underline{T}(\lambda,\cdot) : \Xi \to \Xi$ exist such that $\overline{T}(\lambda,\zeta) \in \overline{F}(\lambda,\zeta)$

and $\underline{T}(\lambda,\zeta) \in \underline{F}(\lambda,\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$.

(iii) The proof is the same as the proof of (ii).

(iv) For each $\zeta \in E$, it follows from the measurability of ℓ and $\mathsf{d}^p(\cdot,\zeta)$ that $F(\zeta) \in \mathscr{B}(\Xi)$. Then using the same argument as in (ii), there exists a $\boldsymbol{\nu}$-measurable selection $T : E \to \Xi$ such that $T(\zeta) \in F(\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta \in E$.

(v) For each $\zeta \in \Xi$, it follows from the measurability of ℓ, M, and $\mathsf{d}^p(\cdot,\zeta)$ that $F(\zeta) \in \mathscr{B}(\Xi)$. Then using the same argument as in (ii), there exists a $\boldsymbol{\nu}$-measurable selection $T : \Xi \to \Xi$ such that $T(\zeta) \in F(\zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. □

Proof of Lemma 2.5. (i) It follows from Definition 2.4 of κ that $\int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) = -\infty$ for all $\lambda < \kappa$ and $\int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)$ is finite for all $\lambda > \kappa$, and thus $h(\lambda) = \infty$ for all $\lambda < \kappa$ and $h(\lambda)$ is finite for all $\lambda > \kappa$.

(ii) It follows from Lemma 2.3(i) that $h(\lambda)$ is the sum of a linear function $\lambda\theta^p$ and an (extended real-valued) convex function $-\int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta)$ on $[0,\infty)$. Thus h is convex.

(iii) Note that (i) and (ii) imply that h is continuous everywhere except possibly at κ. We show that h is lower-semicontinuous at κ. Consider any sequence $\{\lambda_n\}_n$ such that $\lambda_n \downarrow \kappa$. Since $\Phi(\cdot,\zeta)$ is upper-semicontinuous for all $\zeta \in \Xi$, it follows that $\Phi(\kappa,\zeta) \geq \limsup_{n\to\infty} \Phi(\lambda_n,\zeta)$. Also, $\Phi(\lambda,\zeta) \leq -\ell(\zeta)$ for all λ and ζ. Thus $\liminf_{n\to\infty}[-\Phi(\lambda_n,\zeta)] - \ell(\zeta) \geq -\Phi(\kappa,\zeta) - \ell(\zeta) \geq 0$ for all ζ. Hence it follows from Fatou's lemma that

$$\begin{aligned}
\liminf_{n\to\infty} h(\lambda_n) - \int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) &= \liminf_{n\to\infty} \left\{ \lambda_n\theta^p + \int_\Xi [-\Phi(\lambda_n,\zeta) - \ell(\zeta)]\boldsymbol{\nu}(d\zeta) \right\} \\
&\geq \kappa\theta^p + \int_\Xi \liminf_{n\to\infty} [-\Phi(\lambda_n,\zeta) - \ell(\zeta)]\boldsymbol{\nu}(d\zeta) \\
&\geq \kappa\theta^p + \int_\Xi [-\Phi(\kappa,\zeta) - \ell(\zeta)]\boldsymbol{\nu}(d\zeta) \\
&= h(\kappa) - \int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta)
\end{aligned}$$

Since $\left|\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta)\right| < \infty$, it follows that $\liminf_{n\to\infty} h(\lambda_n) \geq h(\kappa)$, and thus h is lower-semicontinuous.

(iv) Since $\Phi(\lambda, \zeta) \leq -\ell(\zeta)$, it follows that $h(\lambda) \geq \lambda\theta^p + \int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) \to \infty$ as $\lambda \to \infty$.

(v) The result follows from (i)–(iv). □

Proof of Lemma A.1. Note that if $x = 0$, then the inequality holds for any $C_p(\varepsilon) \geq 1$. Next we consider the case with $x > 0$, and we let $t := y/x$. Let

$$t_0(\varepsilon) \;:=\; \sup\{t > 0 \;:\; 1 + \varepsilon \geq (1+t)^p\}.$$

Note that $t_0(\varepsilon) > 0$. Next let

$$C_p(\varepsilon) \;:=\; \max\left\{1,\; \sup_{t \geq t_0(\varepsilon)} \frac{(1+t)^{p-1}}{t^{p-1}}\right\}.$$

Note that $C_p(\varepsilon) < \infty$ because $\lim_{t\to\infty}(1+t)^{p-1}/t^{p-1} = 1$. Next, consider

$$f(t) \;:=\; 1 + \varepsilon + C_p(\varepsilon)t^p - (1+t)^p$$

Note that $f(t) \geq 0$ for all $t \in [0, t_0(\varepsilon)]$. Also, $f'(t) = C_p(\varepsilon)pt^{p-1} - p(1+t)^{p-1} \geq 0$ for all $t \in [t_0(\varepsilon), \infty)$. Therefore $f(t) \geq 0$ for all $t \geq 0$, which establishes the inequality for $x > 0$. □

Proof of Lemma A.2. It follows from Lemma A.1 with $\varepsilon := \frac{\lambda - \lambda_1}{2\lambda_1}$ that

$$\lambda_1 \mathsf{d}^p(\xi, \zeta^0) \;\leq\; \frac{\lambda + \lambda_1}{2}\mathsf{d}^p(\xi, \zeta) + \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta, \zeta^0)$$

for all $\xi, \zeta, \zeta^0 \in \Xi$. Thus

$$\begin{aligned}
\lambda \mathsf{d}^p(\xi,\zeta) - \ell(\xi) &= \frac{\lambda-\lambda_1}{2}\mathsf{d}^p(\xi,\zeta) - \ell(\xi) + \frac{\lambda+\lambda_1}{2}\mathsf{d}^p(\xi,\zeta) \\
&\geq \frac{\lambda-\lambda_1}{2}\mathsf{d}^p(\xi,\zeta) - \ell(\xi) + \lambda_1\mathsf{d}^p(\xi,\zeta^0) - \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta,\zeta^0) \\
&\geq \frac{\lambda-\lambda_1}{2}\mathsf{d}^p(\xi,\zeta) + \Phi(\lambda_1,\zeta^0) - \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta,\zeta^0).
\end{aligned}$$

Hence, for every $\xi \in \Xi$ that satisfies $\lambda\mathsf{d}^p(\xi,\zeta) - \ell(\xi) < \Phi(\lambda,\zeta) + \delta$ for some $\delta \geq 0$, it holds that

$$\begin{aligned}
&\frac{\lambda-\lambda_1}{2}\mathsf{d}^p(\xi,\zeta) < \Phi(\lambda,\zeta) - \Phi(\lambda_1,\zeta^0) + \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta,\zeta^0) + \delta \\
\Rightarrow\ &\frac{\lambda-\lambda_1}{2}\limsup_{\delta\downarrow 0}\Big\{\sup_{\xi\in\Xi}\big\{\mathsf{d}^p(\xi,\zeta) \ :\ \lambda\mathsf{d}^p(\xi,\zeta) - \ell(\xi) < \Phi(\lambda,\zeta)+\delta\big\}\Big\} \\
&\qquad\qquad \leq \limsup_{\delta\downarrow 0}\Big\{\Phi(\lambda,\zeta) - \Phi(\lambda_1,\zeta^0) + \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta,\zeta^0) + \delta\Big\} \\
\Rightarrow\ &\frac{\lambda-\lambda_1}{2}\overline{D}(\lambda,\zeta) \leq \Phi(\lambda,\zeta) - \Phi(\lambda_1,\zeta^0) + \lambda_1 C_p(\varepsilon)\mathsf{d}^p(\zeta,\zeta^0)
\end{aligned}$$

□

Proof of Lemma A.4. Since Ξ is separable, ∂C has a countable dense subset $\{\xi^i\}_{i=1}^\infty$. For each ξ^i, there exists $\xi'^i \in \Xi\backslash\mathrm{cl}(\Xi)$ such that $\varepsilon_i := 2\mathsf{d}(\xi^i,\xi'^i) < \varepsilon$. Thus $\partial C = \bigcup_{i=1}^\infty B_{\varepsilon_i}(\xi^i)$, where $B_{\varepsilon_i}(\xi^i)$ is the open ball centered at ξ^i with radius ε_i. Define

$$i^*(\xi) := \min_{i\geq 0}\{i : \xi \in B_{\varepsilon_i}(\xi^i)\}, \quad \xi \in \partial C,$$

and

$$T_\varepsilon(\xi) := \xi_{i^*(\xi)}, \quad \xi \in \partial C.$$

Then T_ε satisfies the requirements in the lemma. □

Proof of Lemma A.3. The proof is a simple exercise in Calculus. Using condition (A.20),

we obtain the bound estimations

$$||F_1(x,y)-F_1(x',y)||_{X,*} \leq \frac{1}{n}\sum_{i=1}^{n} ||\partial_x \ell(x,y^i+\widehat{\xi^i})-\partial_x \ell(x',y^i+\widehat{\xi^i})||_{X,*} \leq L_{11}||x-x'||_X+M_{11},$$

$$\begin{aligned}
||F_1(x',y)-F_1(x',y')||_{X,*} &\leq \frac{1}{n}\sum_{i=1}^{n} ||\partial_x \ell(x',y^i+\widehat{\xi^i})-\partial_x \ell(x',y'^i+\widehat{\xi^i})||_{X,*} \\
&\leq \frac{1}{n}\sum_{i=1}^{n} L_{12}||y^i-y'^i||_\Xi + M_{12} \\
&\leq \frac{L_{12}}{n}\left(\sum_{i=1}^{n}||y^i-y'^i||_\Xi^p\right)^{1/p} n^{1/q} + M_{12} \\
&= L_{12}||y-y'||_Y + M_{12},
\end{aligned}$$

$$\begin{aligned}
||F_2(x,y)-F_2(x',y)||_{Y,*} &= n^{-1}\left(\sum_{i=1}^{n}||\partial_\xi \ell(x,y^i+\widehat{\xi^i})-\partial_\xi \ell(x',y^i+\widehat{\xi^i})||_\Xi^q\right)^{1/q} \\
&\leq n^{-1}\sum_{i=1}^{n}[L_{21}||x-x'||_X + M_{21}] \\
&= L_{21}||x-x'||_X + M_{21},
\end{aligned}$$

$$\begin{aligned}
||F_2(x',y)-F_2(x',y')||_{Y,*} &= n^{-1}\left(\sum_{i=1}^{n}||\partial_\xi \ell(x',y^i+\widehat{\xi^i})-\partial_\xi \ell(x',y'^i+\widehat{\xi^i})||_\Xi^q\right)^{1/q} \\
&\leq n^{-1}\left(\sum_{i=1}^{n}[L_{22}||y^i-y'^i||_\Xi + M_{22}]^q\right)^{1/q} \\
&\leq n^{-1}L_{22}\left(\sum_{i=1}^{n}||y^i-y'^i||_\Xi^q\right)^{1/q} + M_{22} \\
&\leq N^{\max(1/q-1/p,0)-1}L_{22}||y-y'||_Y + M_{22}.
\end{aligned}$$

Combining the above inequalities we have

$$
\begin{aligned}
||F(z) - F(z')||_{Z,*} &= \sqrt{||F_1(z) - F_1(z')||^2_{X,*} + ||F_2(z) - F_2(z')||^2_{Y,*}} \\
&\leq 2\max\{||F_1(z) - F_1(z')||_{X,*}, ||F_2(z) - F_2(z')||_{Y,*}\} \\
&\leq 4\max\{L_{11}||x - x'||_X + M_{11}, L_{12}||y - y'||_Y + M_{12}, \\
&\quad L_{21}||x - x'||_X + M_{21}, N^{\max(1/q-1/p,0)-1}L_{22}||y - y'||_Y + M_{22}\}.
\end{aligned}
$$

□

A.2.2 Proofs of Corollaries

Proof of Corollary 2.1.

(1) Recall from Lemma 2.3(i) that there is a set $B \in \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$ such that $\boldsymbol{\nu}(B) = 1$, and $\Phi(\lambda, \zeta) > -\infty$ for all $\lambda > \kappa$ and all $\zeta \in B$. Note that if ℓ is upper-semicontinuous, and bounded subsets of (Ξ, d) are totally bounded, then for $\delta = 0$, it holds that $\underline{F}(\lambda, \zeta)$ and $\overline{F}(\lambda, \zeta)$ in Lemma 2.4(iii) are nonempty for all $\lambda > \kappa$ and all $\zeta \in B$. Next we show that $\overline{D}_0(\cdot, \zeta)$ and $\underline{D}_0(\cdot, \zeta)$ are nonincreasing. Consider any $\lambda_2 > \lambda_1$ and any $\zeta \in \Xi$ such that $\Phi(\lambda_1, \zeta) > -\infty$. Consider any $\xi_i \in \Xi$ such that $\lambda_i \mathsf{d}^p(\xi_i, \zeta) - \ell(\xi_i) = \Phi(\lambda_i, \zeta)$ for $i = 1, 2$. Then it follows as in the proof of Lemma 2.3(i) that $\mathsf{d}^p(\xi_2, \zeta) \leq \mathsf{d}^p(\xi_1, \zeta)$. Therefore $\overline{D}_0(\lambda_2, \zeta) \leq \underline{D}_0(\lambda_1, \zeta) \leq \overline{D}_0(\lambda_1, \zeta)$.

Next we show that, for all $\zeta \in B$, it holds that $\overline{D}_0(\cdot, \zeta)$ is upper-semicontinuous and $\underline{D}_0(\cdot, \zeta)$ is lower-semicontinuous at all $\lambda > \kappa$. Consider any $\lambda > \kappa$ and any sequence $\{\lambda_n\}_n$ such that $\lambda_n \to \lambda$ as $n \to \infty$ and $\lambda_n \in \big((\lambda + \kappa)/2, \lambda + \delta\big)$ for all n, for some $\delta > 0$. For each n and each $\zeta \in B$, consider any $\xi^n \in \arg\min_{\xi \in \Xi} \big\{\lambda_n \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\big\}$. Note that $\mathsf{d}^p(\xi^n, \zeta) \in \big[\overline{D}_0(\lambda + \delta, \zeta), \underline{D}_0\big((\lambda + \kappa)/2, \zeta\big)\big]$ for all n. Since bounded subsets of (Ξ, d) are totally bounded, it is sufficient to consider subsequences of $\{\xi^n\}_n$ that converge to some $\xi^* \in \Xi$. It follows from the upper-semicontinuity of ℓ and the continuity of $\Phi(\cdot, \zeta)$ at all

$\lambda > \kappa$ that

$$\lambda \mathsf{d}^p(\xi^*, \zeta) - \ell(\xi^*) \;\leq\; \liminf_{n\to\infty} \{\lambda_n \mathsf{d}^p(\xi^n, \zeta) - \ell(\xi^n)\} \;=\; \liminf_{n\to\infty} \Phi(\lambda_n, \zeta) \;=\; \Phi(\lambda, \zeta)$$

and thus $\xi^* \in \arg\min_{\xi\in\Xi} \left\{\lambda \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\right\}$. Since $\underline{D}_0(\lambda, \zeta) \leq \mathsf{d}^p(\xi^*, \zeta) = \lim_{n\to\infty} \mathsf{d}^p(\xi^n, \zeta) \leq \overline{D}_0(\lambda, \zeta)$, it follows that $\overline{D}_0(\cdot, \zeta)$ is upper-semicontinuous and $\underline{D}_0(\cdot, \zeta)$ is lower-semicontinuous at all $\lambda > \kappa$ for all $\zeta \in B$.

Next we show that for all $\lambda > \kappa$ and all $\zeta \in B$, it holds that $\partial\Phi(\lambda, \zeta)/\partial\lambda- = \overline{D}_0(\lambda, \zeta)$ and $\partial\Phi(\lambda, \zeta)/\partial\lambda+ = \underline{D}_0(\lambda, \zeta)$. Consider any $\zeta \in B$ and any $\lambda_2 > \lambda_1 > \kappa$. Consider any $\xi^i \in \arg\min_{\xi\in\Xi} \left\{\lambda_i \mathsf{d}^p(\xi, \zeta) - \ell(\xi)\right\}$ for $i = 1, 2$. Then it follows as in the proof of Lemma 2.3(iii) that

$$\mathsf{d}^p(\xi^2, \zeta) \;\leq\; \frac{\Phi(\lambda_2, \zeta) - \Phi(\lambda_1, \zeta)}{\lambda_2 - \lambda_1} \;\leq\; \mathsf{d}^p(\xi^1, \zeta).$$

Then it follows from the definitions of $\overline{D}_0$ and $\underline{D}_0$ that

$$\overline{D}_0(\lambda_2, \zeta) \;\leq\; \frac{\Phi(\lambda_2, \zeta) - \Phi(\lambda_1, \zeta)}{\lambda_2 - \lambda_1} \;\leq\; \underline{D}_0(\lambda_1, \zeta).$$

Setting $\lambda_2 = \lambda$ and letting $\lambda_1 \uparrow \lambda$, it follows from the upper-semicontinuity of $\overline{D}_0(\cdot, \zeta)$ that

$$\overline{D}_0(\lambda, \zeta) \;\leq\; \frac{\partial}{\partial\lambda-}\Phi(\lambda, \zeta) \;\leq\; \lim_{\lambda_1\uparrow\lambda} \underline{D}_0(\lambda_1, \zeta) \;\leq\; \lim_{\lambda_1\uparrow\lambda} \overline{D}_0(\lambda_1, \zeta) \;\leq\; \overline{D}_0(\lambda, \zeta)$$

and hence

$$\frac{\partial}{\partial\lambda-}\Phi(\lambda, \zeta) \;=\; \overline{D}_0(\lambda, \zeta)$$

Similarly, setting $\lambda_1 = \lambda$ and letting $\lambda_2 \downarrow \lambda$, it follows from the lower-semicontinuity of $\underline{D}_0(\cdot, \zeta)$ that

$$\underline{D}_0(\lambda, \zeta) \;\leq\; \lim_{\lambda_2\downarrow\lambda} \underline{D}_0(\lambda_2, \zeta) \;\leq\; \lim_{\lambda_2\downarrow\lambda} \overline{D}_0(\lambda_2, \zeta) \;\leq\; \frac{\partial}{\partial\lambda+}\Phi(\lambda, \zeta) \;\leq\; \underline{D}_0(\lambda, \zeta)$$

and hence

$$\frac{\partial}{\partial \lambda +}\Phi(\lambda,\zeta) \;=\; \underline{D}_0(\lambda,\zeta)$$

Next we show that if condition (i) or (ii) or (iii) holds, then there exists a primal optimal distribution. First suppose that condition (i) holds: there exists a dual minimizer $\lambda^* > \kappa$. Since for $\delta = 0$, it holds that $\underline{F}(\lambda^*,\zeta)$ and $\overline{F}(\lambda^*,\zeta)$ in Lemma 2.4(iii) are nonempty for all $\zeta \in B$, it follows that there exists $\boldsymbol{\nu}$-measurable mappings $\overline{T}, \underline{T} : \Xi \to \Xi$ such that

$$\begin{aligned}
\overline{T}(\zeta) \;&\in\; \Big\{\xi \in \Xi \;:\; \lambda^* \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \;=\; \Phi(\lambda^*,\zeta),\; \mathsf{d}^p(\xi,\zeta) \;=\; \overline{D}_0(\lambda^*,\zeta)\Big\},\\
\underline{T}(\zeta) \;&\in\; \Big\{\xi \in \Xi \;:\; \lambda^* \mathsf{d}^p(\xi,\zeta) - \ell(\xi) \;=\; \Phi(\lambda^*,\zeta),\; \mathsf{d}^p(\xi,\zeta) \;=\; \underline{D}_0(\lambda^*,\zeta)\Big\}
\end{aligned}$$

for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. As in the proof of Theorem 2.1, it follows from the first-order optimality conditions $\frac{\partial}{\partial \lambda -}h(\lambda^*) \le 0$ and $\frac{\partial}{\partial \lambda +}h(\lambda^*) \ge 0$ that

$$\begin{aligned}
\theta^p \;&\ge\; \frac{\partial}{\partial \lambda +}\left(\int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_\Xi \frac{\partial}{\partial \lambda +}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) && \text{(A.2)}\\
&=\; \int_\Xi \underline{D}_0(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \int_\Xi \mathsf{d}^p(\underline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) && \text{(A.3)}\\
\theta^p \;&\le\; \frac{\partial}{\partial \lambda -}\left(\int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta)\right) \;=\; \int_\Xi \frac{\partial}{\partial \lambda -}\Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) && \text{(A.4)}\\
&=\; \int_\Xi \overline{D}_0(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \int_\Xi \mathsf{d}^p(\overline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta).
\end{aligned}$$

Let $q \in [0,1]$ be such that

$$q\int_\Xi \mathsf{d}^p(\underline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + (1-q)\int_\Xi \mathsf{d}^p(\overline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \theta^p.$$

Let

$$\boldsymbol{\mu}^* \;:=\; q\underline{T}_\#\boldsymbol{\nu} + (1-q)\overline{T}_\#\boldsymbol{\nu}.$$

Then

$$\mathcal{W}_p^p(\boldsymbol{\mu}^*, \boldsymbol{\nu}) \;\leq\; q\int_\Xi \mathsf{d}^p(\underline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) + (1-q)\int_\Xi \mathsf{d}^p(\overline{T}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \theta^p$$

and thus $\boldsymbol{\mu}^*$ is primal feasible. Also,

$$\begin{aligned}
&\int_\Xi \ell(\xi)\boldsymbol{\mu}^*(d\xi)\\
&= q\int_\Xi \ell(\underline{T}(\zeta))\boldsymbol{\nu}(d\zeta) + (1-q)\int_\Xi \ell(\overline{T}(\zeta))\boldsymbol{\nu}(d\zeta)\\
&= q\int_\Xi \big[\lambda^*\mathsf{d}^p(\underline{T}(\zeta),\zeta) - \Phi(\lambda^*,\zeta)\big]\boldsymbol{\nu}(d\zeta) + (1-q)\int_\Xi \big[\lambda^*\mathsf{d}^p(\overline{T}(\zeta),\zeta) - \Phi(\lambda^*,\zeta)\big]\boldsymbol{\nu}(d\zeta)\\
&= \lambda^*\theta^p - \int_\Xi \Phi(\lambda^*,\zeta)\boldsymbol{\nu}(d\zeta) \;=\; v_D.
\end{aligned}$$

Therefore $\boldsymbol{\mu}^*$ is primal optimal.

Suppose that condition (ii) holds: $\lambda^* = \kappa > 0$ is the unique dual minimizer, $\boldsymbol{\nu}\big(\{\zeta \in \Xi : \arg\min_{\xi\in\Xi}\{\kappa\mathsf{d}^p(\xi,\zeta) - \ell(\xi)\} = \varnothing\}\big) = 0$, and

$$\int_\Xi \underline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \theta^p \;\leq\; \int_\Xi \overline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta).$$

Then it follows in the same way as in the proof for condition (i) that there exists a primal optimal distribution.

Suppose that condition (iii) holds: $\lambda^* = \kappa = 0$ is the unique dual minimizer, $\arg\max_{\xi\in\Xi}\{\ell(\xi)\}$ is nonempty, and

$$\int_\Xi \underline{D}_0(0,\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \theta^p.$$

Then, for $\delta = 0$, the sets $\underline{F}(\lambda^*,\zeta)$ in Lemma 2.4(iii) are given by

$$\begin{aligned}
\underline{F}(\lambda^*,\zeta) &= \Big\{\xi \in \Xi \,:\, -\ell(\xi) \,=\, \Phi(\lambda^*,\zeta),\; \mathsf{d}^p(\xi,\zeta) \,\leq\, \underline{D}_0(\lambda^*,\zeta)\Big\}\\
&= \Big\{\xi \in \arg\max_{\xi\in\Xi}\{\ell(\xi)\} \,:\, \mathsf{d}^p(\xi,\zeta) = \underline{D}_0(0,\zeta)\Big\}
\end{aligned}$$

and are non-empty for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. Thus there exists a $\boldsymbol{\nu}$-measurable mapping $\underline{T} : \Xi \to \Xi$ such that $\underline{T}(\zeta) \in \underline{F}(\lambda^*, \zeta)$ for $\boldsymbol{\nu}$-almost all $\zeta \in \Xi$. Let $\boldsymbol{\mu}^* := \underline{T}_{\#}\boldsymbol{\nu}$. Then

$$\mathcal{W}_p^p(\boldsymbol{\mu}^*, \boldsymbol{\nu}) \;\leq\; \int_\Xi \mathsf{d}^p(\underline{T}(\zeta), \zeta)\boldsymbol{\nu}(d\zeta) \;=\; \int_\Xi \underline{D}_0(0, \zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \theta^p$$

and thus $\boldsymbol{\mu}^*$ is primal feasible. Furthermore,

$$\int_\Xi \ell(\xi)\boldsymbol{\mu}^*(d\xi) \;=\; \int_\Xi \ell(\underline{T}(\zeta))\boldsymbol{\nu}(d\zeta) \;=\; \max_{\xi\in\Xi} \ell(\xi) \;=\; v_D$$

and thus $\boldsymbol{\mu}^*$ is primal optimal. Therefore we have shown that if condition (i) or (ii) or (iii) holds, then there exists a primal optimal distribution.

Next we show that if there exists a primal optimal distribution, then condition (i) or (ii) or (iii) holds. Consider any primal feasible distribution $\boldsymbol{\mu}$. Let $\boldsymbol{\gamma} \in \mathcal{P}(\Xi \times \Xi)$ denote the corresponding optimal solution in definition (4.1) of Wasserstein distance $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$, and let $\boldsymbol{\gamma}_\zeta$ denote the corresponding conditional distribution of ξ given ζ. Since $\boldsymbol{\mu}$ is feasible, it holds that $\int_\Xi \int_\Xi \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \leq \theta^p$. Lemma 2.5(v) established existence of a dual minimizer $\lambda^* \in [\kappa, \infty)$. Note that

$$\begin{aligned}
& v_D - \int_\Xi \ell(\xi)\boldsymbol{\mu}(d\xi) \\
= \;& \left[\lambda^*\theta^p - \int_\Xi \Phi(\lambda^*, \zeta)\boldsymbol{\nu}(d\zeta)\right] \\
& - \left[\int_{\Xi^2} [\ell(\xi) - \lambda^*\mathsf{d}^p(\xi, \zeta)]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) + \int_{\Xi^2} \lambda^*\mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right] \\
= \;& \lambda^*\left[\theta^p - \int_\Xi \int_\Xi \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right] \\
& + \left[\int_\Xi \int_\Xi [\lambda^*\mathsf{d}^p(\xi, \zeta) - \ell(\xi) - \Phi(\lambda^*, \zeta)]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right]
\end{aligned}$$

For $\boldsymbol{\mu}$ to be primal optimal, it must hold that $v_D - \int_\Xi \ell(\xi)\boldsymbol{\mu}(d\xi) = 0$. Since $\lambda^* \geq 0$, $\theta^p - \int_\Xi \int_\Xi \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \geq 0$, and $\lambda^*\mathsf{d}^p(\xi, \zeta) - \ell(\xi) - \Phi(\lambda^*, \zeta) \geq 0$ for all (ξ, ζ), it follows that all of the following must hold for $\boldsymbol{\mu}$ to be primal optimal:

(a) $\lambda^* \left[\theta^p - \int_\Xi \int_\Xi \mathsf{d}^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right] = 0$.

(b) $\int_\Xi [\lambda^* \mathsf{d}^p(\xi,\zeta) - \ell(\xi) - \Phi(\lambda^*,\zeta)]\boldsymbol{\gamma}_\zeta(d\xi) = 0$ for $\boldsymbol{\nu}$-almost all ζ, which in turn implies that $\arg\min_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\zeta) - \ell(\xi)\} \neq \varnothing$ for $\boldsymbol{\nu}$-almost all ζ, and the conditional distribution $\boldsymbol{\gamma}_\zeta$ should be supported on $\arg\min_{\xi\in\Xi}\{\lambda^*\mathsf{d}^p(\xi,\zeta) - \ell(\xi)\}$ for $\boldsymbol{\nu}$-almost all ζ.

Next we show that these conditions imply that condition (i) or (ii) or (iii) holds. Since there is a dual minimizer $\lambda^* \in [\kappa, \infty)$, one of the following conditions must hold:

(1) There is a dual minimizer $\lambda^* > \kappa$.

(2) The unique dual minimizer satisfies $\lambda^* = \kappa > 0$.

(3) The unique dual minimizer satisfies $\lambda^* = \kappa = 0$.

If (1) holds, then condition (i) holds, and the proof is complete.

Next suppose that (2) holds, and that $\boldsymbol{\mu}$ is a primal optimal solution. Condition (b) implies that $\boldsymbol{\nu}\big(\{\zeta \in \Xi : \arg\min_{\xi\in\Xi}\{\kappa\mathsf{d}^p(\xi,\zeta) - \ell(\xi)\} = \varnothing\}\big) = 0$. Next we show that

$$\int_\Xi \underline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \theta^p \;\leq\; \int_\Xi \overline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta).$$

It follows from (a) that

$$\theta^p \;=\; \int_\Xi \int_\Xi \mathsf{d}^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \;\leq\; \int_\Xi \overline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta)$$

If $\theta^p < \int_\Xi \underline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta)$, then it follows as in (A.2) that

$$\frac{\partial}{\partial\lambda+}\int_\Xi \Phi(\kappa,\zeta)\boldsymbol{\nu}(d\zeta) \;=\; \int_\Xi \underline{D}_0(\kappa,\zeta)\boldsymbol{\nu}(d\zeta) \;>\; \theta^p.$$

Then there exists a $\lambda > \kappa$ such that $\lambda\theta^p - \int_\Xi \Phi(\lambda,\zeta)\boldsymbol{\nu}(d\zeta) < \kappa\theta^p - \int_\Xi \Phi(\kappa,\zeta)\boldsymbol{\nu}(d\zeta)$, contradicting $\lambda^* = \kappa$ being a dual minimizer. Therefore, if (2) holds, then condition (ii)

holds.

Next suppose that (3) holds. Condition (b) implies that $\arg\max_{\xi\in\Xi}\{\ell(\xi)\} \neq \varnothing$. Suppose that $\boldsymbol{\mu}$ is a primal optimal solution. Then

$$\int_\Xi \underline{D}_0(0,\zeta)\boldsymbol{\nu}(d\zeta) \;\leq\; \int_\Xi\int_\Xi \mathsf{d}^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \;\leq\; \theta^p.$$

Therefore, if (3) holds, then condition (iii) holds.

(2) If $-\ell(\zeta) \leq \inf_{\xi\in\Xi}\{\kappa\mathsf{d}^p(\xi,\zeta)-\ell(\xi)\}$ $\boldsymbol{\nu}$-almost everywhere, i.e., $\ell(\xi)-\ell(\zeta)\leq \kappa\mathsf{d}^p(\xi,\zeta)$, then for any $\lambda>\kappa$, $\Phi(\lambda,\zeta)=\ell(\zeta)$. Hence the dual optimal solution $\lambda^*=\kappa$.

Otherwise there exists a set $E\subset\Xi$ with $\boldsymbol{\nu}(E)>0$ such that $\ell(\zeta)>\Phi(\kappa,\zeta)$ for all $\zeta\in E$, and thus $\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) > \int_\Xi \Phi(\kappa,\zeta)\boldsymbol{\nu}(d\zeta)$. Then by continuity (follows from concavity) of $\int_\Xi \Phi(\cdot,\zeta)\boldsymbol{\nu}(d\zeta)$, there exists $\lambda_0>\kappa$ such that $\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) > \int_\Xi \Phi(\lambda_0,\zeta)\boldsymbol{\nu}(d\zeta)$. For such λ_0, using the upper-semicontinuity and totally boundedness assumptions and Lemma 2.4, there exists a $\boldsymbol{\nu}$-measurable map $T_{\lambda_0}:\Xi\to\Xi$, such that $\lambda_0\mathsf{d}^p(T_{\lambda_0}(\zeta),\zeta)-\ell(T_{\lambda_0}(\zeta))=\Phi(\lambda_0,\zeta)$, and

$$\varepsilon := \int_\Xi \mathsf{d}^p(T_{\lambda_0}(\zeta),\zeta)\boldsymbol{\nu}(d\zeta) > 0,$$

since otherwise $\int_\Xi \ell(\zeta)\boldsymbol{\nu}(d\zeta) = \int_\Xi \Phi(\lambda_0,\zeta)\boldsymbol{\nu}(d\zeta)$. Choose $\theta<\varepsilon^{1/p}$, then we claim that $\lambda=\kappa$ is cannot be optimal. Indeed, according to Case 2 in the proof of Theorem 2.1, $\lambda^*=\kappa$ implies $\int_\Xi \mathsf{d}^p(T^0_\lambda(\zeta),\zeta)\boldsymbol{\nu}(d\zeta)<\theta^p<\varepsilon$ for all $\lambda>\kappa$, and in particular, for $\lambda=\lambda_0$. Thus we arrive at a contradiction.

(3) This directly follows from the fact that in the proof for necessity in (1), in each case we construct an optimal solution with the the structure described as in (3).

(4) For any primal feasible solution $\boldsymbol{\mu}$, let $\boldsymbol{\gamma}^{\boldsymbol{\mu}}$ be a minimizer in the definition (4.1) of $\mathcal{W}_p(\boldsymbol{\mu},\boldsymbol{\nu})$ and let $\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta$ be the conditional distribution of ξ given ζ when the joint distribution

of (ξ, ζ) is $\boldsymbol{\gamma}^{\boldsymbol{\mu}}$. We define $T^{\boldsymbol{\mu}} : \Xi \to \Xi$ by

$$T^{\boldsymbol{\mu}}(\zeta) := \mathbb{E}_{\boldsymbol{\gamma}_\zeta}[\xi].$$

Then it follows from Ξ being convex that $T^{\boldsymbol{\mu}}(\zeta) \in \Xi$ for all $\zeta \in \Xi$. It follows from $\mathsf{d}^p(\cdot, \zeta)$ being convex for all $\zeta \in \Xi$ and Jensen's inequality that

$$\mathcal{W}_p^p(T^{\boldsymbol{\mu}}_{\#}\boldsymbol{\nu}, \boldsymbol{\nu}) \leq \int_\Xi \mathsf{d}^p(\mathbb{E}_{\boldsymbol{\gamma}_\zeta}[\xi], \zeta)\boldsymbol{\nu}(d\zeta) \leq \int_\Xi \mathbb{E}_{\boldsymbol{\gamma}_\zeta}[\mathsf{d}^p(\xi, \zeta)]\boldsymbol{\nu}(d\zeta) = \mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \theta^p.$$

Also, it follows from ℓ being concave that

$$\int_\Xi \ell(T^{\boldsymbol{\mu}}(\zeta))\boldsymbol{\nu}(d\zeta) = \int_\Xi \ell(\mathbb{E}_{\boldsymbol{\gamma}_\zeta}[\xi])\boldsymbol{\nu}(d\zeta) \geq \int_\Xi \mathbb{E}_{\boldsymbol{\gamma}_\zeta}[\ell(\xi)]\boldsymbol{\nu}(d\zeta) = \mathbb{E}_{\boldsymbol{\mu}}[\ell].$$

Hence $T^{\boldsymbol{\mu}}_{\#}\boldsymbol{\nu}$ is an feasible solution with objective no worse than $\boldsymbol{\mu}$, thus the result follows.

□

Proof of Proposition A.1. According to Theorem 4.1 in [75], the results follows if we can provide an upper bound on Θ_Y. Let $h = (h_1, \ldots, h_N) \in \Xi^n$ and $m \leq 2$. The first-order directional derivative of ω_Y at y along h is given by

$$\gamma D\omega_Y(y)[h] = \sum_{i=1}^n ||\xi^i||_{\Xi_i}^{m-2}\langle \xi^i, h_i\rangle.$$

The second-order directional derivative of ω_Y at y along $[h, h]$ is given by

$$\begin{aligned}\gamma D^2\omega_Y(y)[h, h] &= \sum_{i=1}^n ||\xi^i||_{\Xi_i}^{m-2}||h_i||_{\Xi_i}^2 + (m-2)\sum_{i=1}^n ||\xi^i||_{\Xi_i}^{m-4}\langle \xi^i, h_i\rangle^2 \\ &\geq (m-1)\sum_{i=1}^n ||\xi^i||_{\Xi_i}^{m-2}||h_i||_{\Xi_i}^2,\end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality and the condition $m \leq 2$.

On the other hand,

$$||h||_Y^2 = \left(\sum_{i=1}^n ||h_i||_{\Xi_i}^p\right)^{2/p} \leq \left(\sum_{i=1}^n ||h_i||_{\Xi_i}\right)^2 \leq \left(\sum_{i=1}^n ||\xi||_{\Xi_i}^{m-2}||h_i||_{\Xi_i}^2\right)\left(\sum_{i=1}^n ||\xi^i||_{\Xi_i}^{2-m}\right),$$

where the first inequality follows from $p \geq 1$ and the second inequality follows from Cauchy-Schwarz inequality. Combining the above two inequalities yields

$$||h||_Y^2 \leq \frac{\gamma}{m-1} D^2\omega_Y(y)[h,h]\left(\sum_{i=1}^n ||\xi^i||_{\Xi_i}^{2-m}\right). \tag{A.5}$$

Noting that $\sum_{i=1}^n ||\xi^i||_{\Xi_i}^p \leq N\theta^p$, set $t_i = ||\xi^i||_{\Xi_i}^p$, whence $\sum_{i=1}^n t_i \leq N\theta^p$. It follows from Hölder's inequality that

$$\sum_{i=1}^n ||\xi^i||_{\Xi_i}^m \leq \theta^m N^{\max(m/p,1)}. \tag{A.6}$$

Hence $\sum_{i=1}^n ||\xi^i||_{\Xi_i}^{2-m} \leq \theta^{2-m}N^{\max(\frac{2-m}{p},1)}$. Thus in view of (A.5), ω_Y is strongly convex with modulus 1 with respect to $||\cdot||_Y$ if

$$\frac{\gamma}{m-1}\theta^{2-m}N^{\max(\frac{2-m}{p},1)} = 1. \tag{A.7}$$

Now let us compute an upper bound of Θ_Y. According to (A.7), we have

$$\begin{aligned}\Theta_Y = \max_{y\in Y}\omega_Y(y) - \min_{y\in Y}\omega_Y(y) &\leq \frac{1}{m\gamma}\sum_{i=1}^n ||\xi^i||_{\Xi_i}^m \\ &\leq \frac{1}{m(m-1)}\theta^{2-m}N^{\max(\frac{2-m}{p},1)}\left(\sum_{i=1}^n ||\xi^i||_{\Xi_i}^m\right).\end{aligned}$$

Using (A.6) we obtain that

$$\Theta_Y \leq \frac{\theta^2}{m(m-1)}N^{\max(\frac{2-m}{p},1)+\max(\frac{m}{p},1)}.$$

When $p > 1$, choosing $m = \min(2,p)$, we have $\Theta_Y \leq \frac{\theta^2}{m(m-1)}N^2$; when $p = 1$, choosing

$m = 1 + \frac{1}{\ln n}$, we have $\Theta_Y \leq e\theta^2 N^2 \ln n$. It can be easily checked that γ defined in (A.23) satisfies condition (A.7). □

A.2.3 Proofs of Propositions

Proof of Proposition 2.1. Let $\mathcal{Q} := \{\boldsymbol{\mu} \in \mathcal{P}(\Xi) \,:\, \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) < \infty\}$. For any $\boldsymbol{\mu} \in \mathcal{Q}$, let $\boldsymbol{\gamma}^{\boldsymbol{\mu}} \in \mathcal{P}(\Xi^2)$ denote a minimizer in the definition (4.1) of $\mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu})$, and let $\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta$ denote the conditional distribution of ξ given ζ when the joint distribution of (ξ, ζ) is $\boldsymbol{\gamma}^{\boldsymbol{\mu}}$. Then by the tower property of conditional probability it holds that

$$\int_\Xi \ell(\xi)\boldsymbol{\mu}(d\xi) \;=\; \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}(d\xi, d\zeta) = \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta),$$

and

$$\mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) \;=\; \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}^{\boldsymbol{\mu}}(d\xi, d\zeta) \;=\; \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta).$$

Then

$$v_P \;=\; \sup_{\boldsymbol{\mu}\in\mathcal{Q}} \left\{ \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \;:\; \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \leq \theta^p \right\}.$$

Next we show that

$$v_P \;\leq\; \sup_{\boldsymbol{\mu}\in\mathcal{Q}} \inf_{\lambda\geq 0} \left\{ \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) + \lambda \left(\theta^p - \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) \right) \right\}. \quad \text{(A.8)}$$

If $\int_\Xi \ell(\xi)\boldsymbol{\mu}(d\xi) < \infty$ for all $\boldsymbol{\mu} \in \mathcal{Q}$, then for any $\boldsymbol{\mu} \in \mathfrak{M} = \{\boldsymbol{\mu} \in \mathcal{P}(\Xi) \,:\, \mathcal{W}_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \theta\}$ it holds that

$$\inf_{\lambda\geq 0} \left\{ \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta) + \lambda\Big(\theta^p - \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\Big) \right\} \;=\; \int_{\Xi^2} \ell(\xi)\boldsymbol{\gamma}^{\boldsymbol{\mu}}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)$$

and for any $\boldsymbol{\mu} \in \mathcal{Q} \setminus \mathfrak{M}$ it holds that

$$\inf_{\lambda \geq 0} \left\{ \int_{\Xi^2} \ell(\xi) \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) + \lambda \Big(\theta^p - \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta) \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) \Big) \right\} \;=\; -\infty$$

Thus the objective functions in (Primal) and the right side of (A.8) are the same for all $\boldsymbol{\mu} \in \mathcal{Q}$, and therefore (A.8) holds as an equality.

Otherwise, if $\int_\Xi \ell(\xi) \boldsymbol{\mu}(d\xi) = \infty$ for some $\boldsymbol{\mu} \in \mathcal{Q}$, then for any $\lambda \geq 0$ it holds that

$$\int_{\Xi^2} \ell(\xi) \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) + \lambda \left(\theta^p - \int_{\Xi^2} \mathsf{d}^p(\xi, \zeta) \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) \right) \;=\; \infty,$$

because $\int_{\Xi^2} \mathsf{d}^p(\xi, \zeta) \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) = \mathcal{W}_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) < \infty$, and thus (A.8) holds. Therefore

$$\begin{aligned} v_P \;&\leq\; \inf_{\lambda \geq 0} \left\{ \lambda \theta^p + \sup_{\boldsymbol{\mu} \in \mathcal{Q}} \left\{ \int_{\Xi^2} \big[\ell(\xi) - \lambda \mathsf{d}^p(\xi, \zeta) \big] \boldsymbol{\gamma}_\zeta^{\boldsymbol{\mu}}(d\xi) \boldsymbol{\nu}(d\zeta) \right\} \right\} \\ &\leq\; \inf_{\lambda \geq 0} \left\{ \lambda \theta^p + \int_\Xi \sup_{\xi \in \Xi} \big[\ell(\xi) - \lambda \mathsf{d}^p(\xi, \zeta) \big] \boldsymbol{\nu}(d\zeta) \right\} \\ &=\; v_D. \end{aligned}$$

□

Proof of Proposition 2.2. If $\kappa = \infty$, then for any $n > 0$ it holds that

$$\phi^n(\zeta) \;:=\; \inf_{\xi \in \Xi} \{ n \mathsf{d}^p(\xi, \zeta) - \ell(\xi) + \ell(\zeta) \} \;\notin\; L^1(\boldsymbol{\nu}).$$

Observe that $\phi^n(\zeta) = \Phi(n, \zeta) + \ell(\zeta)$. Hence, for any $n > 0$, there exists $E^n \in \mathscr{B}_{\boldsymbol{\nu}}$ with $\boldsymbol{\nu}(E^n) > 0$, such that $\phi^n(\zeta) < -n$ for $\boldsymbol{\nu}$-almost all $\zeta \in E^n$. By Lemma 2.4(iv), there exists a $\boldsymbol{\nu}$-measurable mapping $T^n : E^n \to \Xi$ such that

$$T^n(\zeta) \;\in\; F^n(\zeta) \;:=\; \{ \xi \in \Xi \,:\, \ell(\xi) - \ell(\zeta) > n \mathsf{d}^p(\xi, \zeta) + n \}$$

for $\boldsymbol{\nu}$-almost all $\zeta \in E^n$. For $m = 1, 2, \ldots$, consider the set

$$E_m^n \;:=\; \{\zeta \in E^n \;:\; \mathsf{d}^p(T^n(\zeta), \zeta) \le m\}\,.$$

Note that $E_m^n \in \mathscr{B}_{\boldsymbol{\nu}}$ for all m. Then $\lim_{m\to\infty} E_m^n = E^n$ and thus $\lim_{m\to\infty} \boldsymbol{\nu}(E_m^n) = \boldsymbol{\nu}(E^n) > 0$. Hence, for each n, there exists a m^n such that $\boldsymbol{\nu}(E_{m^n}^n) > 0$, and $\int_{E_{m^n}^n} \mathsf{d}^p(T^n(\zeta), \zeta)\boldsymbol{\nu}(d\zeta) \le m^n < \infty$. Let $\overline{T}^n$ be the restriction of T^n on $E_{m^n}^n$. Note that $\overline{T}^n$ is also $\boldsymbol{\nu}$-measurable.

For each n, let

$$p^n \;:=\; \min\left\{1,\; \frac{\theta^p}{\int_{E_{m^n}^n} \mathsf{d}^p(\overline{T}^n(\zeta), \zeta)\boldsymbol{\nu}(d\zeta)}\right\}$$

and define the distribution

$$\boldsymbol{\mu}^n \;:=\; p^n \overline{T}^n_{\#}\boldsymbol{\nu} + (1 - p^n)\boldsymbol{\nu}$$

Then $\boldsymbol{\mu}^n$ is a primal feasible solution, and

$$\int_\Xi \ell d\boldsymbol{\mu}^n - \int_\Xi \ell d\boldsymbol{\nu} \;\ge\; p^n \left(n \int_{E_{m^n}^n} \mathsf{d}^p(\overline{T}^n(\zeta), \zeta)\boldsymbol{\nu}(d\zeta) + n\right) \;\ge\; \min\{n\theta^p, n\}.$$

Since n can be chosen arbitrarily large, it follows that $v_P = \infty = v_D$. □

Proof of Proposition 2.3. Since $-\mathbb{1}\{\xi \in \mathrm{int}(C)\}$ is upper semicontinuous and binary-valued, by Corollary 2.1 the worst-case distribution of $\min_{\boldsymbol{\mu}\in\mathfrak{M}} \boldsymbol{\mu}(\mathrm{int}(C))$ exists. Thus it suffices to show that for any $\varepsilon > 0$, there exists $\boldsymbol{\mu} \in \mathfrak{M}$ such that $\boldsymbol{\mu}(C) \le \min_{\boldsymbol{\mu}\in\mathfrak{M}} \boldsymbol{\mu}(\mathrm{int}(C)) + \varepsilon$. Observe that there exists an optimal transportation plan $\boldsymbol{\gamma}_0$ such that

$$\mathrm{supp}\, \boldsymbol{\gamma}_0 \subset \big(\mathrm{supp}\, \boldsymbol{\nu} \times \mathrm{supp}\, \boldsymbol{\nu}\big) \cup \big((\mathrm{supp}\, \boldsymbol{\nu} \cap \mathrm{int}(C)) \times \partial C\big).$$

Set $\boldsymbol{\mu}_0 := \pi^2_{\#}\boldsymbol{\gamma}_0$, then $\boldsymbol{\mu}_0$ is an optimal solution for $\min_{\boldsymbol{\mu}\in\mathfrak{M}} \boldsymbol{\mu}(\mathrm{int}(C))$.

If $\boldsymbol{\mu}_0(\partial C) = 0$, there is nothing to show, so we assume that $\boldsymbol{\mu}_0(\partial C) > 0$. We first consider the case $\boldsymbol{\nu}(\mathrm{int}(C)) = 0$ (and thus $\boldsymbol{\mu}_0$ can be chosen to be $\boldsymbol{\nu}$ and the worst-case value is 0). By Lemma A.4, we can define a Borel map T_ε which maps each $\xi \in \partial C$ to some $\xi' \in \Xi \setminus \mathrm{cl}(C)$ with $d(\xi, \xi') < \varepsilon \in (0, \theta)$ and is an identity mapping elsewhere. We further define a distribution $\boldsymbol{\mu}_\varepsilon$ by

$$\boldsymbol{\mu}_\varepsilon(A) := \boldsymbol{\mu}_0(A \setminus \partial C) + \boldsymbol{\mu}_0\{\xi \in \partial C : T_\varepsilon(\xi) \in A\}, \text{ for all Borel set } A \subset \Xi.$$

Then $\mathcal{W}_p(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\mu}_0) = \mathcal{W}_p(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\nu}) \le \varepsilon < \theta$ and $\boldsymbol{\mu}_\varepsilon(C) = \boldsymbol{\mu}_0(\mathrm{int}(C))$.

Now let us consider $\boldsymbol{\nu}(\mathrm{int}(C)) > 0$. For any $\varepsilon \in (0, \theta)$, we define a distribution $\boldsymbol{\mu}'_\varepsilon$ by

$$\begin{aligned}\boldsymbol{\mu}'_\varepsilon(A) :=& \boldsymbol{\mu}_0(A \cap \mathrm{int}(C)) + \frac{\varepsilon}{\theta}\Big(\boldsymbol{\gamma}_0\{(A \cap \mathrm{int}(C)) \times \partial C\} + \boldsymbol{\nu}(A \cap \partial C)\Big) \\ &+ \Big(1 - \frac{\varepsilon}{\theta}\Big)\boldsymbol{\mu}_0\{\xi \in \partial C : T_\varepsilon(\xi) \in A\} + \boldsymbol{\mu}_0(A \setminus \mathrm{cl}(C)), \quad \text{for all Borel set } A \subset \Xi.\end{aligned}$$

Then

$$\begin{aligned}\boldsymbol{\mu}'_\varepsilon(C) &= \boldsymbol{\mu}_0(\mathrm{int}(C)) + \frac{\varepsilon}{\theta}[\boldsymbol{\mu}_0(\partial C) - \boldsymbol{\nu}(\partial C) + \boldsymbol{\nu}(\partial C)] + 0 + 0 \\ &\le \boldsymbol{\mu}_0(\mathrm{int}(C)) + \frac{\varepsilon}{\theta}.\end{aligned}$$

Note that $\mathcal{W}_p^p(\boldsymbol{\mu}_0, \boldsymbol{\nu}) = \int_{\mathrm{int}(C)\times\partial C} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_0(d\xi, d\zeta)$, it follows that

$$\begin{aligned}\mathcal{W}_p(\boldsymbol{\mu}'_\varepsilon, \boldsymbol{\nu}) &\le \Big(1 - \frac{\varepsilon}{\theta}\Big)\int_{\mathrm{int}(C)\times\partial C} \mathsf{d}^p(\xi, \zeta)\boldsymbol{\gamma}_0(d\xi, d\zeta) + \Big(1 - \frac{\varepsilon}{\theta}\Big)\varepsilon + 0 \\ &\le \theta.\end{aligned}$$

Hence the proof is completed. □

Proof of Proposition 2.4. Observe that

$$\begin{aligned}\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}[\boldsymbol{\eta}(\beta^{-1}(1))] &= \inf_{\boldsymbol{\mu}\in\mathcal{P}(\Xi)} \big\{\mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}[\boldsymbol{\eta}(\beta^{-1}(1))] : \min_{\boldsymbol{\gamma}\in\Gamma(\boldsymbol{\mu},\boldsymbol{\nu})} \mathbb{E}_{\boldsymbol{\gamma}}[\mathsf{d}(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})] \le \theta\big\} \\ &= \inf_{\boldsymbol{\gamma}\in\mathcal{P}(\Xi^2)} \big\{\mathbb{E}_{(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\sim\boldsymbol{\gamma}}[\boldsymbol{\eta}(\beta^{-1}(1))] :\ \mathbb{E}_{\boldsymbol{\gamma}}[\mathsf{d}(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})] \le \theta,\ \pi^2_{\#}\boldsymbol{\gamma} = \boldsymbol{\nu}\big\}.\end{aligned} \tag{A.9}$$

For any $\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)$, denote by $\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}$ the conditional distribution of $\bar{\theta} := \mathsf{d}(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})$ given $\widehat{\boldsymbol{\eta}}$, and by $\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}$ the conditional distribution of $\boldsymbol{\eta}$ given $\widehat{\boldsymbol{\eta}}$ and $\bar{\theta}$. Using tower property of conditional probability, we have that for any $\boldsymbol{\gamma} \in \mathcal{P}(\Xi^2)$ with $\pi^2_{\#}\boldsymbol{\gamma} = \boldsymbol{\nu}$,

$$\mathbb{E}_{(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})\sim\boldsymbol{\gamma}}[\boldsymbol{\eta}(\beta^{-1}(1))] = \mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\Big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}\big[\mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))]\big]\Big],$$

and

$$\mathbb{E}_{\boldsymbol{\gamma}}[\mathsf{d}(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})] = \mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}[\bar{\theta}]\big].$$

Observe that the right-hand side of the second equation above does not depend on $\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}$. Thereby (A.9) can be reformulated as

$$\begin{aligned}&\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\mu}}[\boldsymbol{\eta}(\beta^{-1}(1))] \\ &= \inf_{\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}\}_{\widehat{\boldsymbol{\eta}}},\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}\}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}} \Big\{\mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\Big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}\big[\mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))]\big]\Big] : \mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}[\bar{\theta}]\big] \le \theta\Big\} \\ &= \inf_{\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}\}_{\widehat{\boldsymbol{\eta}}}} \left\{\mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\left[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}\left[\inf_{\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}\}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))]\right]\right] : \mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}[\bar{\theta}]\big] \le \theta\right\},\end{aligned} \tag{A.10}$$

where the second equality follows from interchangeability principle (cf. Theorem 14.60 in [143]). We claim that

$$\begin{aligned}&\inf_{\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}\}_{\widehat{\boldsymbol{\eta}}}} \left\{\mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\left[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}\left[\inf_{\{\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}\}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}} \mathbb{E}_{\boldsymbol{\eta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))]\right]\right] : \mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}[\bar{\theta}]\big] \le \theta\right\} \\ &= \inf_{\rho\in\mathcal{P}(\mathcal{B}([0,1])\times\Xi)} \Big\{\mathbb{E}_{(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\sim\rho}[\tilde{\boldsymbol{\eta}}(\mathrm{int}(\beta^{-1}(1)))] :\ \mathbb{E}_{(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\sim\rho}\big[W_1(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\big] \le \theta,\ \pi^2_{\#}\rho = \boldsymbol{\nu}\Big\}.\end{aligned} \tag{A.11}$$

Indeed, let ρ be any feasible solution of the right-hand side of (A.11). We denote by $\rho_{\widehat{\boldsymbol{\eta}}}$ the conditional distribution of $\bar{\bar{\theta}} := W_1(\tilde{\boldsymbol{\eta}}, \widehat{\boldsymbol{\eta}})$ given $\widehat{\boldsymbol{\eta}}$ and by $\rho_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}$ the conditional distribution of $\tilde{\boldsymbol{\eta}}$ given $\widehat{\boldsymbol{\eta}}$ and $\bar{\bar{\theta}}$. When $\widehat{\boldsymbol{\eta}} = 0$ (i.e. no arrival) or $\bar{\bar{\theta}} = 0$, set $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}}} = \delta_0$ and $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} = \widehat{\boldsymbol{\eta}}$, that is, we choose $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}}}$ and $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}$ be such that $\boldsymbol{\eta} = \widehat{\boldsymbol{\eta}}$. When $\widehat{\boldsymbol{\eta}} \neq 0$ and $\bar{\bar{\theta}} > 0$, applying Corollary 2.1 (Example 2.7) and Proposition 2.3 to the problem $\min_{\tilde{\boldsymbol{\eta}} \in \mathcal{B}([0,1])}\{\tilde{\boldsymbol{\eta}}(\beta^{-1}(1)) : W_1(\tilde{\boldsymbol{\eta}}, \widehat{\boldsymbol{\eta}}) \leq \bar{\bar{\theta}}\}$, we have that for any $\varepsilon > 0$, there exists an ε-optimal solution $\tilde{\boldsymbol{\eta}}$ of the form

$$\tilde{\boldsymbol{\eta}} = \sum_{\substack{i=1 \\ i \neq i_0}}^{\widehat{\boldsymbol{\eta}}([0,1])} \delta_{\tilde{\xi}^i} + p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} \delta_{\tilde{\xi}^+_{i_0}} + (1 - p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}) \delta_{\tilde{\xi}^-_{i_0}},$$

where $1 \leq i_0 \leq \widehat{\boldsymbol{\eta}}([0,1])$, $p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} \in [0,1]$, and $\tilde{\xi}_i \in [0,1]$ for all $i \neq i_0$ and $\tilde{\xi}^{\pm}_{i_0} \in [0,1]$. Define

$$\boldsymbol{\eta}^{\pm}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} := \sum_{\substack{i=1 \\ i \neq i_0}}^{\widehat{\boldsymbol{\eta}}([0,1])} \delta_{\tilde{\xi}^i} + \delta_{\tilde{\xi}^{\pm}_{i_0}}.$$

It follows that $\boldsymbol{\eta}^{\pm}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}([0,1]) = \widehat{\boldsymbol{\eta}}([0,1])$, and

$$p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} \boldsymbol{\eta}^{+}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}(\beta^{-1}(1)) + (1 - p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}) \boldsymbol{\eta}^{-}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}(\beta^{-1}(1)) \leq \varepsilon + \min_{\tilde{\boldsymbol{\eta}} \in \mathcal{B}([0,1])} \left\{ \tilde{\boldsymbol{\eta}}(\mathrm{int}(\beta^{-1}(1))) : W_1(\tilde{\boldsymbol{\eta}}, \widehat{\boldsymbol{\eta}}) \leq \bar{\bar{\theta}} \right\}, \tag{A.12}$$

and

$$p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} W_1(\boldsymbol{\eta}^{+}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) + (1 - p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}) W_1(\boldsymbol{\eta}^{-}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) \leq \bar{\bar{\theta}}. \tag{A.13}$$

Define $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}}}$ and $\bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}$ by

$$\begin{aligned} \bar{\boldsymbol{\gamma}}_{\widehat{\boldsymbol{\eta}}}(C) := \int_0^{\infty} \big[& p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}} \mathbb{1}\{W_1(\boldsymbol{\eta}^{+}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) \in C\} \\ & + (1 - p_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}) \mathbb{1}\{W_1(\boldsymbol{\eta}^{-}_{\widehat{\boldsymbol{\eta}},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) \in C\} \big] \rho_{\widehat{\boldsymbol{\eta}}}(d\bar{\bar{\theta}}), \ \ \forall \text{ Borel set } C \subset [0, \infty), \end{aligned}$$

and

$$\bar{\gamma}_{\widehat{\eta},\bar{\theta}}(A) := \int_0^\infty \int_\Xi \Big[p_{\widehat{\eta},\bar{\bar{\theta}}} \mathbb{1}\big\{ \boldsymbol{\eta}^+_{\widehat{\eta},\bar{\bar{\theta}}} \in A,\ \mathcal{W}_1(\boldsymbol{\eta}^+_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) = \bar{\theta} \big\} \\ + (1 - p_{\widehat{\eta},\bar{\bar{\theta}}}) \mathbb{1}\big\{ \boldsymbol{\eta}^-_{\widehat{\eta},\bar{\bar{\theta}}} \in A,\ \mathcal{W}_1(\boldsymbol{\eta}^-_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) = \bar{\theta} \big\} \Big] \rho_{\widehat{\eta},\bar{\bar{\theta}}}(d\boldsymbol{\eta}) \rho_{\widehat{\eta}}(d\bar{\bar{\theta}}),$$

for all Borel set $A \subset \Xi$. Then $(\{\bar{\gamma}_{\widehat{\eta}}\}_{\widehat{\eta}}, \{\bar{\gamma}_{\widehat{\eta},\bar{\theta}}\}_{\widehat{\eta},\bar{\theta}})$ is a feasible solution to the left-hand side of (A.11). Indeed, by condition (ii), we have $\mathsf{d}(\boldsymbol{\eta}^\pm_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) = W_1(\boldsymbol{\eta}^\pm_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}})$, hence (A.13) implies that $p_{\widehat{\eta},\bar{\bar{\theta}}} \mathsf{d}(\boldsymbol{\eta}^+_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) + (1 - p_{\widehat{\eta},\bar{\bar{\theta}}}) \mathsf{d}(\boldsymbol{\eta}^-_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) \le \bar{\bar{\theta}}$. Then taking expectation on both sides,

$$\begin{aligned} \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\bar{\gamma}_{\widehat{\eta}}}[\bar{\theta}]\big] &= \int_\Xi \int_0^\infty \Big[p_{\widehat{\eta},\bar{\bar{\theta}}} d(\boldsymbol{\eta}^+_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) + (1 - p_{\widehat{\eta},\bar{\bar{\theta}}}) d(\boldsymbol{\eta}^-_{\widehat{\eta},\bar{\bar{\theta}}}, \widehat{\boldsymbol{\eta}}) \Big] \rho_{\widehat{\eta}}(d\bar{\bar{\theta}}) \boldsymbol{\nu}(d\widehat{\boldsymbol{\eta}}) \\ &= \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}[\bar{\theta}]\big] \le \theta, \end{aligned}$$

hence $\{\bar{\gamma}_{\widehat{\eta}}\}_{\widehat{\eta}}$ is feasible. Similarly, taking expectation on both sides of (A.12), we have that $\mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\Big[\mathbb{E}_{\bar{\theta}\sim\bar{\gamma}_{\widehat{\eta}}}\big[\mathbb{E}_{\eta\sim\bar{\gamma}_{\widehat{\eta},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))]\big]\Big] \le \varepsilon + \mathbb{E}_\rho[\tilde{\boldsymbol{\eta}}(\beta^{-1}(1))]$. Let $\varepsilon \to 0$, we obtain that

$$\begin{aligned} &\inf_{\{\gamma_{\widehat{\eta}}\}_{\widehat{\eta}}} \left\{ \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\left[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}\left[\inf_{\{\gamma_{\widehat{\eta},\bar{\theta}}\}_{\widehat{\eta},\bar{\theta}}} \mathbb{E}_{\eta\sim\gamma_{\widehat{\eta},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))] \right] \right] : \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}[\bar{\theta}]\big] \le \theta \right\} \\ \le &\inf_{\rho\in\mathcal{P}(\mathcal{B}([0,1])\times\Xi)} \Big\{ \mathbb{E}_{(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\sim\rho}[\tilde{\boldsymbol{\eta}}(\mathrm{int}(\beta^{-1}(1)))] :\ \mathbb{E}_{(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\sim\rho}\big[W_1(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\big] \le \theta,\ \pi^2_\# \rho = \boldsymbol{\nu} \Big\}. \end{aligned}$$

To show the opposite direction of the above inequality, observe that $\inf_{\boldsymbol{\mu}_{\widehat{\eta},\bar{\theta}}} \mathbb{E}_{\eta\sim\boldsymbol{\mu}_{\widehat{\eta},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))] = \inf_{\boldsymbol{\eta}\in\Xi}\{\boldsymbol{\eta}(\beta^{-1}(1)) : d(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}}) = \bar{\theta}\}$. Hence

$$\begin{aligned} &\inf_{\{\gamma_{\widehat{\eta}}\}_{\widehat{\eta}}} \left\{ \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\left[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}\left[\inf_{\{\gamma_{\widehat{\eta},\bar{\theta}}\}_{\widehat{\eta},\bar{\theta}}} \mathbb{E}_{\eta\sim\gamma_{\widehat{\eta},\bar{\theta}}}[\boldsymbol{\eta}(\beta^{-1}(1))] \right] \right] : \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}[\bar{\theta}]\big] \le \theta \right\} \\ = &\inf_{\{\gamma_{\widehat{\eta}}\}_{\widehat{\eta}}} \left\{ \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\Big[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}\big[\inf_{\boldsymbol{\eta}\in\Xi} \big\{ \boldsymbol{\eta}(\beta^{-1}(1)) : d(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}}) = \bar{\theta} \big\} \big] \Big] : \mathbb{E}_{\widehat{\eta}\sim\boldsymbol{\nu}}\big[\mathbb{E}_{\bar{\theta}\sim\gamma_{\widehat{\eta}}}[\bar{\theta}]\big] \le \theta \right\}. \end{aligned} \tag{A.14}$$

Let $(\{\boldsymbol{\gamma}_{\widehat{\eta}}\}_{\widehat{\eta}}, \{\boldsymbol{\eta}_{\widehat{\eta},\bar{\theta}}\}_{\widehat{\eta},\bar{\theta}})$ be a feasible solution of the right-hand side of (A.14). Then the joint distribution $\bar{\rho} \in \mathcal{P}(\mathcal{B}([0,1]) \times \Xi)$ defined by

$$\bar{\rho}(B) := \int_{\pi^2(B)} \int_0^\infty \mathbb{1}\{\boldsymbol{\eta}_{\widehat{\eta},\bar{\theta}} \in \pi^1(B)\} \boldsymbol{\gamma}_{\widehat{\eta}}(d\bar{\theta}) \boldsymbol{\nu}(d\widehat{\boldsymbol{\eta}}),\ \ \forall \text{ Borel set } B \subset \mathcal{B}([0,1]) \times \Xi$$

is a feasible solution of the right-hand side of (A.11). By condition (iii), we have that

$$\inf_{\boldsymbol{\eta}\in\Xi}\{\boldsymbol{\eta}(\beta^{-1}(1)) \,:\, d(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}})=\bar{\theta}\} \;\geq\; \inf_{\tilde{\boldsymbol{\eta}}\in\mathcal{B}([0,1])}\Big\{\tilde{\boldsymbol{\eta}}(\mathrm{int}(\beta^{-1}(1))) \,:\, W_1(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\leq\bar{\theta}\Big\},$$

and thus $\mathbb{E}_{\widehat{\boldsymbol{\eta}}\sim\boldsymbol{\nu}}\Big[\mathbb{E}_{\bar{\theta}\sim\boldsymbol{\gamma}_{\widehat{\boldsymbol{\eta}}}}\big[\inf_{\boldsymbol{\eta}}\{\boldsymbol{\eta}(\beta^{-1}(1)) \,:\, d(\boldsymbol{\eta},\widehat{\boldsymbol{\eta}}) = \bar{\theta}\}\big]\Big] \geq \mathbb{E}_{(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\sim\rho}[\tilde{\boldsymbol{\eta}}(\mathrm{int}(\beta^{-1}(1)))]$. Therefore we prove the opposite direction and (A.11) holds. Together with (A.10), we obtain (2.23).

It then follows that it suffices to only consider policy β such that $\beta^{-1}(1)$ is an open set. Then by Corollary 2.1 (Example 2.7), the problem $\min_{\tilde{\boldsymbol{\eta}}\in\mathcal{B}([0,1])}\Big\{\tilde{\boldsymbol{\eta}}(\beta^{-1}(1)) : W_1(\tilde{\boldsymbol{\eta}},\widehat{\boldsymbol{\eta}})\leq \bar{\theta}\Big\}$ admits a worst-case distribution $\boldsymbol{\eta}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}$ and let $\lambda_{\widehat{\boldsymbol{\eta}},\bar{\theta}}$ be the associated dual optimizer. Let $\hat{\Xi} := \{\widehat{\xi}^i_m \,:\, i=1,\ldots,N, t=1,\ldots,M_i\}$. We claim that it suffices to further restrict attention to those policies β such that each connected component of $\beta^{-1}(1)$ contains at least one point in $\hat{\Xi}$. Indeed, suppose there exists a connected component C_0 of $\beta^{-1}(1)$ such that $C_0\cap\hat{\Xi}=\varnothing$. Then for every $\zeta\in\mathrm{supp}\,\widehat{\boldsymbol{\eta}}$, $\arg\min_{\xi\in[0,1]}\big[\mathbb{1}_{\beta^{-1}(1)}(\xi)+|\xi-\zeta|\big]\notin C_0$, and thus $\boldsymbol{\eta}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}(\beta^{-1}(1)) = \boldsymbol{\eta}_{\widehat{\boldsymbol{\eta}},\bar{\theta}}(\beta^{-1}(1)\setminus C_0)$. Hence, $x' := \mathbb{1}_{\{\beta^{-1}(1)\setminus C_0\}}$ achieves a higher objective value $v(x')$ than $v(x)$ and so β cannot be optimal. We finally conclude that there exists $\{\underline{\beta}_j,\overline{\beta}_j\}_{j=1}^M$, where $M\leq \mathrm{card}(\hat{\Xi})$, such that (2.24) holds. □

Using Corollary 2.2 and Proposition 2.4, we have that

$$v\Big(\sum_{j=1}^M \mathbb{1}_{[\underline{\beta}_j,\overline{\beta}_j]}\Big) = \min_{\substack{0\leq p^i\leq 1,\\ \boldsymbol{\eta}^i,\tilde{\boldsymbol{\eta}}^i\in\Xi}}\Bigg\{\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^M\Big[-c(\overline{\beta}_j-\underline{\beta}_j)+p^i\boldsymbol{\eta}^i\{[\underline{\beta}_j,\overline{\beta}_j]\}+(1-p^i)\tilde{\boldsymbol{\eta}}^i\{[\underline{\beta}_j,\overline{\beta}_j]\}\Big] : \frac{1}{n}\sum_{i=1}^n\big[p^iW_1(\boldsymbol{\eta}^i,\widehat{\boldsymbol{\eta}}^i)+(1-p^i)W_1(\tilde{\boldsymbol{\eta}}^i,\widehat{\boldsymbol{\eta}}^i)\big]\leq\theta\Bigg\}. \tag{A.15}$$

By the equivalent definition of one-dimensional Wasserstein distance [144], for $\boldsymbol{\eta}^i = \sum_{i=1}^{M_i}\delta_{\xi^i_m}$, we have that $W_1(\boldsymbol{\eta}^i,\widehat{\boldsymbol{\eta}}^i) = \min_\sigma\sum_{m=1}^{M_i}|\xi^i_m-\widehat{\xi}^i_{\sigma(m)}|$, where the minimum is

taken over all M_i-permutations. Hence

$$
\begin{aligned}
& v\Big(\sum_{j=1}^{M} \mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}\Big) - \sum_{j=1}^{M} -c(\overline{\beta}_j - \underline{\beta}_j) \\
= & \min_{\substack{\xi_m^i, \tilde{\xi}_m^i \in [0,1], \\ 0 \le p^i \le 1}} \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} \Big[p^i \sum_{m=1}^{M_i} \sum_{j=1}^{M} \mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}(\xi_m^i) + (1-p^i) \sum_{m=1}^{M_i} \sum_{j=1}^{M} \mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}(\tilde{\xi}_m^i) \Big] : \\
& \qquad\qquad \frac{1}{n} \sum_{i=1}^{n} \Big[p^i \sum_{m=1}^{M_i} |\xi_m^i - \widehat{\xi}_m^i| + (1-p^i) \sum_{m=1}^{M_i} |\tilde{\xi}_m^i - \widehat{\xi}_m^i| \Big] \le \theta \Bigg\}.
\end{aligned} \tag{A.16}
$$

Using Example 2.6, we have that

$$
\begin{aligned}
& v\Big(\sum_{j=1}^{M} \mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}\Big) - \sum_{j=1}^{M} -c(\overline{\beta}_j - \underline{\beta}_j) \\
= & \min \Bigg\{ \frac{1}{n} \sum_{i=1}^{n} \Big(p^i \sum_{m=1}^{M_i} \sum_{j=1}^{M} [\mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}(\widehat{\xi}_m^i) - (\underline{p}_{mj}^i + \overline{p}_{mj}^i)] \\
& \qquad + (1-p^i) \sum_{m=1}^{M_i} \sum_{j=1}^{M} [\mathbb{1}_{[\underline{\beta}_j, \overline{\beta}_j]}(\widehat{\xi}_m^i) - (\underline{p'}_{mj}^i + \overline{p'}_m^i)] \Big) : \\
& \qquad\qquad \frac{1}{n} \sum_{i=1}^{n} \Big(p^i \sum_{m=1}^{M_i} \sum_{j=1}^{M} \big[\underline{p}_{mj}^i |\underline{\beta}_j - \widehat{\xi}_m^i| + \overline{p}_{mj}^i |\overline{\beta}_j - \widehat{\xi}_m^i| \big] \\
& \qquad\qquad\qquad + (1-p^i) \sum_{m=1}^{M_i} \sum_{j=1}^{M} \big[\underline{p'}_{mj}^i |\underline{\beta}_j - \widehat{\xi}_m^i| + \overline{p'}_m^i |\overline{\beta}_j - \widehat{\xi}_m^i| \big] \Big) \le \theta, \\
& \qquad\qquad\qquad \sum_{j=1}^{M} \big(\underline{p}_{mj}^i + \overline{p}_{mj}^i \big) \le 1, \sum_{j=1}^{M} \big(\underline{p'}_{mj}^i + \overline{p'}_{mj}^i \big) \le 1, \forall i, t \Bigg\},
\end{aligned}
$$

where the minimum is taken over all $p^i, \overline{p}_{mj}^i, \underline{p}_{mj}^i, \overline{p'}_{mj}^i, \underline{p'}_{mj}^i \in [0,1]$. Replacing $p^i \overline{p}_{mj}^i + (1-p^i)\overline{p'}_{mj}^i$ by $\overline{p}_{mj}^i$, and $p^i \underline{p}_{mj}^i + (1-p^i)\underline{p'}_{mj}^i$ by $\underline{p}_{mj}^i$, and noticing that at optimality, $\overline{p}_{mj}^i, \underline{p}_{mj}^i > 0$ only if $\widehat{\xi}_m^i \in [x_{j_-}, \overline{\beta}_j]$, and at most one of $\{\overline{p}_{mj}^i, \underline{p}_{mj}^i\}_{i,t,j}$ can be fractional, we obtain the result.

Proof of Proposition 2.6. The dual optimizer of the inner maximization problem of (2.25) is zero when θ is sufficiently large, whence the worst-case value of the inner maximization

problem equals $\sup_{\xi\in[0,T]} -\ln(\beta(\xi))$. Then the overall objective function equals

$$\int_0^T \beta(t)dt + \sup_{\xi\in[0,T]} -\ln(\beta(\xi)).$$

Set $b = \int_0^T \beta(t)dt$. Then due to the second term $\sup_{\xi\in[0,T]} -\ln(\beta(\xi))$, the solution $\tilde{a}(t) = b/T$ yields an objective value no larger than $\beta(t)$. Hence we complete the proof. □

Proof of Proposition 2.7. Define $C_\beta := \{\xi : -\beta^\top \xi < q\}$ for all w. Similar to Example 2.7, there exists a worst-case distribution $\boldsymbol{\mu}^*$ which attains the infimum $\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \mathbb{P}_{\boldsymbol{\mu}}\{-\beta^\top\xi < q\}$ and there exists maps $\underline{T}^*, \overline{T}^*$ such that for each $\zeta \in \operatorname{supp} \boldsymbol{\nu}$, it holds that $\underline{T}^*(\zeta), \overline{T}^*(\zeta) \in \{\zeta\} \cup \arg\min_{\xi\in\Xi\setminus C_\beta} ||\xi - \zeta||_\infty^p$. With this in mind, let $\boldsymbol{\gamma}^*$ be the optimal transport plan between $\boldsymbol{\nu}$ and $\boldsymbol{\mu}^*$, and let

$$t^* := \boldsymbol{\nu}\text{-ess}\sup_{\zeta\in\Xi} \left\{ \min_{\xi\in\Xi\setminus C_\beta} ||\xi - \zeta||_\infty^p : \zeta \neq \underline{T}^*(\zeta) \right\}.$$

So t^* is the longest distance of transportation among all the points that are transported. (We note that infinity is allowed in the definition of t^*, however, as will be shown, this violates the probability bound.) Then $\boldsymbol{\mu}^*$ transports all the points in $\operatorname{supp}\boldsymbol{\nu} \cap \{\xi : q - t^* < -\beta^\top\xi < b\}$, and possibly a fraction of mass $\alpha^* \in [0,1]$ in $\operatorname{supp}\boldsymbol{\nu} \cap \{\xi : -\beta^\top\xi = q - t^*\}$. Also note that by Hölder's inequality, the distance between two hyperplanes $\{\xi : -\beta^\top\xi = s\}$ and $\{\xi : -\beta^\top\xi = s'\}$ equals to $|s - s'|/||\beta||_1 = |s - s'|$. Using this characterization, let us define a probability measure $\boldsymbol{\nu}_\beta$ on $\mathbb{R}$ by

$$\boldsymbol{\nu}_\beta\{(-\infty, s)\} := \boldsymbol{\nu}\{\xi : -\beta^\top\xi < s\},\ \forall s \in \mathbb{R},$$

then using the changing of measure, the total distance of transportation can be computed by

$$\int_{(\Xi\setminus C_\beta)\times C_\beta} \mathsf{d}^p(\xi,\zeta)\boldsymbol{\gamma}^*(d\xi, d\zeta) = \int_{(q-t^*)+}^{q-} (q-s)^p \boldsymbol{\nu}_\beta(ds) + \alpha^* \boldsymbol{\nu}_\beta(\{q - t^*\})t^{*p} \leq \theta^p. \quad \text{(A.17)}$$

On the other hand, using property of marginal expectation and the characterization of $\boldsymbol{\gamma}^*$,

$$\begin{aligned}\boldsymbol{\mu}^*(C_\beta) &= \int_{C_\beta \times \Xi} \boldsymbol{\gamma}^*(d\xi, d\zeta) \\ &= \boldsymbol{\nu}(C_\beta) - \int_{(\Xi \setminus C_\beta)\times C_\beta} \boldsymbol{\gamma}^*(d\xi, d\zeta) \\ &= 1 - \boldsymbol{\nu}_\beta([q,\infty)) - \alpha^* \boldsymbol{\nu}_\beta(\{q - t^*\}) + \boldsymbol{\nu}_\beta\{(q-t^*, q)\} \\ &= 1 - \boldsymbol{\nu}_\beta(q - t^*, \infty) - \alpha^* \boldsymbol{\nu}_\beta(\{q - t^*\}).\end{aligned}$$

Thereby the condition $\inf_{\boldsymbol{\mu}\in\mathfrak{M}} \boldsymbol{\mu}(C_\beta) \geq 1 - \alpha$ is equivalent to

$$\alpha^* \boldsymbol{\nu}_\beta(\{q - t^*\}) + \boldsymbol{\nu}_\beta(q - t^*, \infty) \leq \alpha. \tag{A.18}$$

Now consider the quantity

$$J := \int_{(\mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}])+}^{q-} (q-s)^p \boldsymbol{\nu}_\beta(ds) + \beta_0 \boldsymbol{\nu}_\beta\big(\{\mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]\}\big)(q - \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}])^p - \theta^p.$$

If $J < 0$, due to the monotonicity in t^* of the right-hand side of (A.17), either $q - t^* < \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]$ or $q - t^* = \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]$ and $\alpha^* > \beta_0$. But in both cases (A.18) is violated. On the other hand if $J \geq 0$, again by monotonicity, either $q - t^* > \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]$, or $q - t^* > \mathrm{VaR}_\alpha[-\beta^\top \boldsymbol{\xi}]$ and $\alpha^* \leq \beta_0^*$ and thus (A.18) is satisfied. Therefore we complete the proof. □

A.3 Selecting Radius θ

We mainly use a classical result on Wasserstein distance from [145]. Let $\boldsymbol{\nu}_N$ be the empirical distribution of ξ obtained from the underlying distribution $\boldsymbol{\nu}_0$. In Theorem 1.1 (see also Remark 1.4) of [145], it is shown that $\mathbb{P}\{W_1(\boldsymbol{\nu}_N, \boldsymbol{\nu}_0) > \theta\} \leq C(\theta) e^{-\frac{\lambda}{2} N\theta^2}$ for some constant λ dependent on $\boldsymbol{\nu}_0$, and C dependent on θ. Since their result holds for general distributions, we here simplify it for our purpose and explicitly compute the constants λ

and C. For a more detailed analysis, we refer the reader to Section 2.1 in [145].

Noticing that by assumption $\operatorname{supp} \boldsymbol{\nu}_0 \subset [0, \bar{B}]$, the truncation step in [145] is no longer needed, thus the probability bound (2.12) (see also (2.15)) of [145] is reduced to

$$\mathbb{P}\{W_1(\boldsymbol{\nu}_N, \boldsymbol{\nu}_0) > \theta\} \leq \max\left(8e\frac{\bar{B}}{\delta}, 1\right)^{\backslash(\frac{\delta}{2})} e^{-\frac{\lambda}{8}N(\theta-\delta)^2}$$

for some constant $\lambda > 0$, $\delta \in (0, \theta)$, where e is the natural logarithm, and $\backslash(\frac{\delta}{2})$ is the minimal number of balls need to cover the support of ξ by balls of radius $\delta/2$ and in our case, $\backslash(\frac{\delta}{2}) = \bar{B}/\delta$. Now let us compute λ. By Theorem 1.1 of [145], λ is the constant appeared in the Talagrand inequality

$$W_1(\boldsymbol{\mu}, \boldsymbol{\nu}_0) \leq \sqrt{\frac{2}{\lambda} I_{\phi_{kl}}(\boldsymbol{\mu}, \boldsymbol{\nu}_0)},$$

where the Kullback-Leibler divergence of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\nu}$ is defined by $I_{\phi_{kl}}(\boldsymbol{\mu}, \boldsymbol{\nu}_0) = +\infty$ if $\boldsymbol{\mu}$ is not absolutely continuous with respect to $\boldsymbol{\nu}_0$, otherwise $I_{\phi_{kl}}(\boldsymbol{\mu}, \boldsymbol{\nu}_0) = \int f \log f d\boldsymbol{\nu}_0$, where f is the Radon-Nikodym derivative $d\boldsymbol{\mu}/d\boldsymbol{\nu}_0$. Corollary 4 in [146] shows that λ can be chosen as

$$\lambda = \left[\inf_{\zeta^0 \in \Xi, \alpha > 0} \frac{1}{\alpha}\left(1 + \log \int e^{\alpha d^2(\xi, \zeta^0)} \boldsymbol{\nu}(d\xi)\right)\right]^{-1},$$

which can be estimated from data. Finally, we obtain a concentration inequality

$$\mathbb{P}\{W_1(\boldsymbol{\nu}_N, \boldsymbol{\nu}_0) > \theta\} \ \leq \ \max\left(8e\frac{\bar{B}}{\delta}, 1\right)^{\frac{\bar{B}}{\delta}} e^{-\frac{\lambda}{8}N(\theta-\delta)^2}. \tag{A.19}$$

In the numerical experiment, we choose δ to make the right-hand side of (A.19) as small as possible, and θ is chosen such that the right-hand side of (A.19) is equal to 0.05.

A.4 Mirror-Prox algorithm for solving Example 2.5

In the following, we briefly describe and set up the algorithm. For a detailed description, we refer the reader to [75]. For ease of notation, set $y := (\xi^1, \ldots, \xi^n)$. We assume that Ξ is a separable Hilbert space such with the metric d induced from some inner product $\langle \cdot, \cdot \rangle$. Set Ξ_i to be the translated space of Ξ under translation mapping $\xi \mapsto \xi - \widehat{\xi^i}$, then a natural norm on Ξ_i is given by

$$||\xi||_{\Xi_i} \ := \ d(\xi, \widehat{\xi^i}), \quad \forall \xi \in \Xi_i.$$

On the product space $\Xi^n := \prod_{i=1}^n \Xi_i$, we define a norm $||\cdot||_Y$ by

$$||y||_Y := ||(\xi^1, \ldots, \xi^n)||_Y := \left(\sum_{i=1}^n ||\xi^i||_{\Xi_i}^p \right)^{1/p}.$$

We introduce the distance generating function

$$\omega_Y(y) = \omega_Y(\xi^1, \ldots, \xi^n) := \frac{1}{m\gamma} \sum_{i=1}^n ||\xi^i||_{\Xi_i}^m,$$

where m, γ are chosen later such that ω_Y is strongly convex with modulus 1 with respect to $||\cdot||_Y$. We also assume that there exists a norm $||\cdot||_X$ on X and a distance generating function $\omega_X(\cdot)$ which is continuous and strongly convex with modulus 1 with respect to $||\cdot||_X$, and admits a continuous selection $\omega'(x)$ of subgradients. Let $\Theta_X := \sup_{x \in X} \omega_X(x) - \inf_{x \in X} \omega_X(x)$.

On the product space $Z \times Y$, we define a norm

$$||z||_Z := \sqrt{\frac{1}{\Theta_X^2} ||x||_X^2 + \frac{1}{\Theta_Y^2} ||y||_Y^2}$$

for any $z \in Z$. It can be easily checked that

$$||z||_{Z,*} = \sqrt{\Theta_X^2 ||x||_{X,*}^2 + \Theta_Y^2 ||y||_{Y,*}^2}$$

defines the dual norm. Suppose there exists $L_{11}, L_{12}, L_{21}, L_{22}, M_{11}, M_{12}, M_{21}, M_{22} \geq 0$ such that for any $x, x' \in X$, $\xi, \xi' \in \Xi$,

$$\begin{cases} ||\partial_x \ell(x,\xi) - \partial_x \ell(x',\xi)||_{X,*} \leq L_{11}||x - x'||_X + M_{11}, \\ ||\partial_x \ell(x,\xi) - \partial_x \ell(x,\xi')||_{X,*} \leq L_{12}||\xi - \xi'||_\Xi + M_{12}, \\ ||\partial_\xi \ell(x,\xi) - \partial_\xi \ell(x',\xi)||_{Y,*} \leq L_{21}||x - x'||_X + M_{21}, \\ ||\partial_\xi \ell(x,\xi) - \partial_\xi \ell(x',\xi)||_{\Xi,*} \leq L_{22}||\xi - \xi'||_\Xi + M_{22}. \end{cases} \tag{A.20}$$

Set

$$\begin{aligned} L &:= \sqrt{2\Theta_X^4 L_{11}^2 + 2\Theta_X^2 \Theta_Y^2 (L_{12}^2 + L_{21}^2) + 2\Theta_Y^4 L_{22}^2}, \\ M &:= \sqrt{2\Theta_X^2 M_{11}^2 + 2\Theta_Y^2 M_{21}^2} + \sqrt{2\Theta_X^2 M_{21}^2 + 2\Theta_Y^2 M_{22}^2}. \end{aligned} \tag{A.21}$$

Define the vector field

$$F(z) := \frac{1}{n}\left[\sum_{i=1}^n \nabla_x \ell(x, \xi^i); \{-\nabla_\xi \ell(x, \xi^i)\}_{i=1}^n\right], \; z \in X \times Y. \tag{A.22}$$

It follows that from Lemma A.3 in the Appendix that

$$||F(z) - F(z')||_{Z,*} \leq L||z - z'||_Z + M.$$

Set

$$\omega(z) := \frac{1}{\Theta_X^2}\omega_X(x) + \frac{1}{\Theta_Y^2}\omega_Y(y),$$

then ω is a distance generating function compatible to $||\cdot||_Z$, and $\Theta := \sup_{z \in Z} \omega(z) - \inf_{z \in Z} \omega(z) = 1$.

Suppose there is a first-order oracle which computes $F(z)$ on each call. The accuracy of a candidate solution $(x, y) \in X \times Y$ is characterized by

$$\varepsilon_{\text{sad}}(x, y) := \max_{(\xi'^1, \ldots, \xi'^n) \in Y} \frac{1}{n}\sum_{i=1}^n \ell(x, \xi'^i) - \min_{x' \in X} \frac{1}{n}\sum_{i=1}^n \ell(x', \xi^i).$$

Given n and θ, the lower bound on oracle complexity of (2.5) using first-order methods is $O(M/\sqrt{t})$ when $L = 0$ and $O(L/t)$ when $M = 0$, and the Mirror-Prox algorithm can achieve the lower bound up to a constant. The Mirror-Prox algorithm is shown in Algorithm 3.

Algorithm 3 Mirror-Prox Algorithm

1: $z_1 := (x_1, y_1) \in X \times Y$.
2: $w_t = \mathrm{Prox}(\gamma_t F(z_t))$, $z_{t+1} = \mathrm{Prox}(\gamma_t F(w_t))$, where $\mathrm{Prox}_z(\gamma_t F(z_t)) := \arg\min_{z \in Z}[V(z, z_t) + \langle \gamma_t F(z_t), z\rangle]$ and $V(z, z') := \omega(z) - \omega(z') - \langle \nabla\omega(z'), z - z'\rangle$.
3: $z^t = \frac{\sum_{\tau=1}^t \gamma_\tau w_\tau}{\sum_{\tau=1}^t \gamma_\tau}$.

Proposition A.1. *Assume (A.20) holds and let L, M be defined in (A.21). Set*

$$m = \begin{cases} 1 + \frac{1}{\ln n}, & p = 1 \\ p, & 1 < p \le 2 \\ 2, & p > 2 \end{cases}, \quad \gamma = \begin{cases} \frac{n}{\ln n}\theta^{\frac{1}{\ln n}-1}, & p = 1 \\ (p-1)\theta^{2-p} n^{2/p-1}, & 1 < p \le 2 \\ n^{2/p-1}, & p > 2 \end{cases}. \tag{A.23}$$

Let γ_τ, $1 \le \tau \le t$ satisfies

$$\gamma_\tau \langle F(z_\tau) - F(w_\tau), w_\tau - z_{\tau+1}\rangle - V(z_\tau, w_\tau) - V(w_\tau, z_{\tau+1}) \le 0,$$

which is definitely satisfied when $\gamma_\tau \in (0, \frac{1}{\sqrt{2}L}]$, or $\gamma_\tau \in (0, 1/L]$ when $M = 0$. Then it holds that

$$\varepsilon_{\mathrm{sad}}(z^t) \le \frac{1 + M^2 \sum_{\tau=1}^t \gamma_\tau^2}{\sum_{\tau=1}^t \gamma_\tau}, \tag{A.24}$$

and

$$\Theta_Y \le \begin{cases} \frac{e\theta^2 N^2 \ln n}{1 + 1/\ln n}, & p = 1 \\ \frac{\theta^2}{p(p-1)} N^2, & 1 < p \le 2, \\ \frac{\theta^2}{2} N^2, & p > 2. \end{cases} \tag{A.25}$$

Note that if $\gamma_\tau = O(1/L)$, by (A.21)(A.24) we have $\varepsilon_{\mathrm{sad}}(z^t) = O(\Theta_Y^2)$. Thus according to (A.25), $\varepsilon_{\mathrm{sad}}(z^t) = O(\theta^4 n^4)$ when $p > 1$ and $\varepsilon_{\mathrm{sad}}(z^t) = O(\theta^4 n^4 \ln^2 n)$ when $p = 1$. We

argue that the complexity of the algorithm may not be very large as it appears. For one thing, we only use DRSO formulation when n is not very large, since otherwise the Sample Average Approximation method provides a relatively accurate solution. For another thing, the radius θ goes to zero as n goes to infinity. Therefore, we believe that the Mirror-Prox algorithm (Algorithm 3) is useful in many practical applications.

APPENDIX B

APPENDIX FOR CHAPTER 3

Lemma B.1. *Let $\kappa > 0$ and $p \geq \kappa + 1$. Then for any $\delta, t > 0$, it holds that*

$$t^{\kappa+1} \;\leq\; \frac{p-1-\kappa}{p-1} \cdot \delta \cdot t \;+\; \frac{\kappa}{p-1} \cdot \delta^{-\frac{p-1-\kappa}{\kappa}} \cdot t^p.$$

Proof of Lemma B.1. When $p = \kappa + 1$, the inequality holds as equality. When $p > \kappa + 1$, set $u = \frac{p-1}{p-1-\kappa}$, $v = \frac{p-1}{\kappa}$. It follows that $\frac{1}{u} + \frac{1}{v} = 1$. Then the result is a consequence of the following Young's inequality

$$\frac{(\delta^{1/u} t^{1/u})^u}{u} + \frac{(\delta^{-1/u} t^{p/v})^v}{v} \;\geq\; t^{\kappa+1}.$$

□

APPENDIX C

APPENDIX FOR CHAPTER 4

Proof of Proposition 4.1. Given a copula $\boldsymbol{\mathcal{C}}$, set $\boldsymbol{\mathcal{C}}_u$ to be the marginal distribution of $\boldsymbol{v}$ given $\boldsymbol{u} = u$. Let $\mathcal{U}$ be the uniform distribution on $[0, 1]$. Under the above condition on d, we have that

$$\omega_{1,\mathsf{d}}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \int_0^1 \mathcal{W}_1(\boldsymbol{\mathcal{C}}_u, \mathcal{U}) du.$$

Then the result follows from the formula for one-dimensional Wasserstein distance [144]. □

Proof of Proposition 4.2. Observe that when choosing ℓ_1-norm, the optimal transportation defining $\mathcal{W}_1(\boldsymbol{\mathcal{C}}^M, \Pi)$ can be chosen such that each point is transported only vertically. Then the computational of Wasserstein distance is reduced to the case in Proposition 1, and thus the result follows. □

Proof of Lemma 4.1. Observe that for any random vector $(\boldsymbol{\xi}, \boldsymbol{\zeta})$ with joint distribution $\boldsymbol{\gamma} \in \mathcal{P}(\Xi \times \Xi)$ and marginals $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}(\Xi)$, it holds that

$$\int_\Xi \ell(\xi) \boldsymbol{\mu}(d\xi) = \int_{\Xi \times \Xi} \ell(\xi) \boldsymbol{\gamma}(d\xi, d\zeta) = \int_\Xi \int_\Xi \ell(\xi) \boldsymbol{\gamma}_\zeta(d\xi) \boldsymbol{\nu}(d\zeta),$$

where $\boldsymbol{\gamma}_\zeta$ represents the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\zeta} = \zeta$. Also note that $\boldsymbol{\mu}$ has marginal F_k if and only if $\int_\Xi f_k(\xi_k) \boldsymbol{\mu}(d\xi) = \int_{\Xi_k} f_k(t) F_k(dt)$ for all $f_k \in B(\Xi_k)$. With the

observations above, using Lagrangian weak duality, we have that

$$
\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathfrak{M}}\left\{\int_{\Xi}\ell(\xi)\boldsymbol{\mu}(d\xi)\right\}\\
&=\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\inf_{\substack{\lambda\geq 0\\ f_k\in\bar{B}(\Xi_k)}}\left\{\int_{\Xi^2}\ell(\xi)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)+\lambda\rho^p-\lambda\int_{\Xi^2}\mathsf{d}_F^p(\xi,\zeta)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right.\\
&\qquad\qquad\left.+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)-\int_{\Xi^2}\sum_k f_k(\xi_k)\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}\\
&\leq\inf_{\substack{\lambda\geq 0\\ f_k\in\bar{B}(\Xi_k)}}\left\{\lambda\rho^p+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)\right.\\
&\qquad\left.+\sup_{\{\boldsymbol{\gamma}_\zeta\}_\zeta\subset\mathcal{P}(\Xi)}\int_{\Xi^2}\big[\ell(\xi)-\sum_k f_k(\xi_k)-\lambda\mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\gamma}_\zeta(d\xi)\boldsymbol{\nu}(d\zeta)\right\}\\
&\leq\inf_{\substack{\lambda\geq 0\\ f_k\in\bar{B}(\Xi_k)}}\left\{\lambda\rho^p+\sum_k\int_{\Xi_k}f_k(t)F_k(dt)\right.\\
&\qquad\left.+\int_{\Xi}\sup_{\xi\in\Xi}\big[\ell(\xi)-\sum_k f_k(\xi_k)-\lambda\mathsf{d}_F^p(\xi,\zeta)\big]\boldsymbol{\nu}(d\zeta)\right\}.
\end{aligned}
$$

□

Proof of Lemma 4.2. We claim that there exists $M>0$ such that

$$
\begin{aligned}
&v_D\\
&=\inf_{\substack{0\leq\lambda\leq M\\ f_k\in\bar{B}(\Xi_k)}}\left\{\lambda\rho^p+\sum_{k=1}^K\int_{\Xi_k}f_k(t)F_k(dt)+\int_{\Xi}\sup_{\xi\in\Xi}\Big[\ell(\xi)-\sum_{k=1}^K f_k(\xi_k)-\lambda\mathsf{d}_F^p(\xi,\zeta)\Big]\boldsymbol{\nu}(d\zeta)\right\}.
\end{aligned}
\tag{C.1}
$$

Indeed, according to the assumption on ℓ, there exists $M>0$ such that $\ell(\xi)\leq M$ for all $\xi\in\Xi$ and thus by choosing $\alpha=\lambda=0$ and $f_k\equiv 0$, we obtain that $v_D\leq M$. On the other hand, fixing $f_k\equiv 0$, the dual objective tends to infinity as $\lambda\to\infty$. Hence the claim holds.

For any feasible solution $(\lambda,\{f_k\}_k)$ of (C.1) such that the dual objective is finite, we are going to define a modification $(\lambda,\{\bar{f}_k\}_k)$ which yields a dual objective value no worse than $(\lambda,\{f_k\}_k)$, but also has a nicer continuity property. The technique used here is the

convexification trick. Setting

$$\Phi(\lambda,\zeta) := \sup_{\xi\in\Xi}\Big\{\ell(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\},$$

we define

$$\bar{f}_1(\xi_1) := \sup_{\substack{\zeta\in\Xi\\ 0\le u_k\le K, k\ge 2}}\Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{k\ge 2} f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\},$$

and inductively define $\bar{f}_k$ by

$$\bar{f}_k(\xi_k) := \sup_{\substack{\zeta\in\Xi\\ 0\le\xi_j\le K, j\ne k}}\Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j<k}\bar{f}_j(\xi_j) - \sum_{j>k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\},\ \forall 2\le k\le K.$$

The definition of Φ implies that

$$f_1(\xi_1) \ge \ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j\ge 2} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta), \quad \forall \xi,\zeta\in\Xi,$$

hence $f_1 \ge \bar{f}_1$. Similarly, the definition of $\bar{f}_{k-1}$ implies that

$$f_k(\xi_k) \ge \ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j<k}\bar{f}_j(\xi_j) - \sum_{j>k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta), \quad \forall \xi,\zeta\in\Xi,$$

hence $f_k \ge \bar{f}_k$ for $k\ge 2$. Hence, for all $1\le k\le K$ it holds that

$$\begin{aligned}\bar{f}_k(\xi_k) &= \sup_{\substack{\zeta\in\Xi\\ 0\le\xi_j\le K, j\ne k}}\Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j<k}\bar{f}_j(\xi_j) - \sum_{j>k} f_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\}\\ &\le \sup_{\substack{\zeta\in\Xi\\ 0\le\xi_j\le K, j\ne k}}\Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j\ne k}\bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\}.\end{aligned}$$

But the definition of $\bar{f}_K$ gives that

$$\bar{f}_k(\xi_k) \geq \sup_{\substack{\zeta\in\Xi \\ 0\leq\xi_j\leq K, j\neq k}} \Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j\neq k} \bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\}, \quad \forall k.$$

Combining the previous two inequalities yields

$$\bar{f}_k(\xi_k) = \sup_{\substack{\zeta\in\Xi \\ 0\leq\xi_j\leq K, j\neq k}} \Big\{\ell(\xi) - \Phi(\lambda,\zeta) - \sum_{j\neq k} \bar{f}_j(\xi_j) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\}, \quad \forall k,$$

which also implies that

$$\sup_{\xi\in\Xi}\Big\{\ell(\xi) - \sum_{k=1}^K \bar{f}_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\} = \Phi(\lambda,\zeta) = \sup_{\xi\in\Xi}\Big\{\ell(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta)\Big\},$$

Together with $\bar{f}_k \leq f_k$, we conclude that $(\lambda, \{\bar{f}_k\}_k)$ yields a dual objective no greater than $(\lambda, \{f_k\}_k)$.

Moreover, since the dual objective remains unchanged if $\{\bar{f}_k\}_k$ is modified into $\{\bar{f}_k + a_k\}_k$, where $a_k \in \mathbb{R}$, so we may assume that $\min_{\Xi_k} \bar{f}_k = 0$. It then follows that $\{f_k\}_k$ are also upper bounded. In addition, the Lipschitz continuity of $\ell - \lambda \mathsf{d}_F^p$ implies $\bar{f}_k$ are also Lipschitz continuous and the Lipschitz constant only depends on that of $\ell - \lambda \mathsf{d}_F^p$.

Now let $(\lambda^{(m)}, \{f_k^{(m)}\}_k)_m$ be a minimizing sequence of (C.1). Using the convexification trick as above, we obtain a sequence $(\lambda^{(m)}, \{\bar{f}_k^{(m)}\}_k)_m$. Then the analysis above implies that $(\bar{f}_k^{(m)})_m$ are uniformly bounded and equi-continuous. Hence by Bolzano-Weierstrass theorem and Arzela-Ascoli theorem, there exists a convergent subsequence. Denote its limit by $(\lambda^*, \{f_k^*\}_{k=1}^K)$. Then by dominate convergence $(\lambda^*, \{f_k^*\}_{k=1}^K)$ is a dual minimizer.

□

Proof of Theorem 4.2. Let us first relax the continuity assumption made in Step 2. We will relax the compactness assumption in the last step. Note that any upper semi-continuous

function satisfying the growth rate condition can be written as the infimum of a non-increasing sequence of Lipschitz continuous functions, for example, by Moreau-Yosida approximation [107]. Thus we can approximate ℓ by a non-increasing sequence of Lipschitz continuous functions ℓ_n and approximate d_F by a non-decreasing sequence of Lipschitz continuous functions d_n. Let us define

$$
\begin{aligned}
v_P^n &:= \sup_{\boldsymbol{\mu}\in\mathfrak{M}} \int_\Xi \ell_n d\boldsymbol{\mu}, \quad v_P^0 := \sup_{\boldsymbol{\mu}\in\mathfrak{M}} \int_\Xi \ell d\boldsymbol{\mu}, \\
v_D^n &:= \inf_{\substack{\lambda\geq 0 \\ f_k\in B(\Xi_k)}} \Big\{ \lambda\rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) \\
&\qquad\qquad + \int_\Xi \sup_{\xi\in\Xi} \Big[\ell_n(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_n^q(\xi,\zeta) \Big] \boldsymbol{\nu}(d\zeta) \Big\}, \\
v_D^0 &:= \inf_{\substack{\lambda\geq 0 \\ f_k\in B(\Xi_k)}} \Big\{ \lambda\rho^p + \sum_{k=1}^K \int_{\Xi_k} f_k(t) F_k(dt) \\
&\qquad\qquad + \int_\Xi \sup_{\xi\in\Xi} \Big[\ell(\xi) - \sum_{k=1}^K f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \Big] \boldsymbol{\nu}(d\zeta) \Big\}.
\end{aligned}
$$

Since $\ell_n \geq \ell$ and $\mathsf{d}_n \leq \mathsf{d}_F$, we have $v_D^0 \leq v_D^n$. From previous steps we know $v_D^n = v_P^n$. In view of $v_D^0 \geq v_P^0$, it remains to show $\lim_{n\to\infty} v_D^n \leq v_D^0$. From Step 4, we know that there exists $\boldsymbol{\mu}_n^*$ such that $\int_\Xi \ell_n d\boldsymbol{\mu}_n^* = v_D^n$. Observe that $\mathfrak{M}$ is tight, then by Prokhorov's theorem, it is relatively compact with respect to the weak topology, and thus $\{\boldsymbol{\mu}^n\}_n$ admits a convergent subsequence, whose limit is denoted by $\boldsymbol{\mu}_0^*$. Then $\int_\Xi \ell_n d\boldsymbol{\mu}_n^* \leq \int_\Xi \ell_m d\boldsymbol{\mu}_n^*$ for all $n \geq m$ implies that

$$
\lim_{n\to\infty} \int_\Xi \ell_n d\boldsymbol{\mu}_n^* \leq \liminf_{n\to\infty} \int_\Xi \ell_m d\boldsymbol{\mu}_n^* \leq \int_\Xi \ell_m d\boldsymbol{\mu}_0^*.
$$

Let $m \to \infty$, by monotone convergence $\lim_{n\to\infty} \int_\Xi \ell_n d\boldsymbol{\mu}_n^* \leq \int_\Xi \ell d\boldsymbol{\mu}_0^* \leq v_P^0$, which concludes the proof.

We next consider the setting where Ξ is not compact. For any $\epsilon > 0$, let $\Xi^\epsilon \subset \Xi$ be a

compact set such that $\boldsymbol{\nu}(\Xi \setminus \Xi^\epsilon) \leq \epsilon$. Set

$$\boldsymbol{\nu}^\epsilon := \frac{\mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}}{\boldsymbol{\nu}(\Xi^\epsilon)},$$

and let F_k^ϵ be the marginal distribution of $\boldsymbol{\nu}^\epsilon$, and Ξ_k^ϵ be its support. Then the previous steps imply that

$$\begin{aligned}
&\sup_{\boldsymbol{\mu}\in\mathcal{P}(\Xi^\epsilon)} \left\{ \int_{\Xi^\epsilon} \ell d\boldsymbol{\mu} :\ W_p(\boldsymbol{\mu}, \boldsymbol{\nu}^\epsilon) \leq \rho,\ \pi_\#^k \boldsymbol{\mu} = \boldsymbol{\nu}_k^\epsilon, \forall k \right\} \\
= &\inf_{\substack{\lambda\geq 0 \\ f_k \in \overline{B}(\Xi_k^\epsilon)}} \left\{ \lambda\rho^p + \sum_k \int_{\Xi_k} f_k F_k^\epsilon \right. \\
&\qquad \left. + \int_\Xi \sup_{\xi\in\Xi^\epsilon} \left[\ell(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi,\zeta) \right] \boldsymbol{\nu}(d\zeta) \right\} \\
=: &v^\epsilon.
\end{aligned}$$

Observe that for any feasible solution $(\lambda, \{f_k\}_k)$ of the dual problem above, the growth condition on ℓ implies that there exists sufficiently large M such that $(\lambda, \{f_k + M\mathbb{1}_{\Xi\setminus\Xi^\epsilon}\}_k)$ is a feasible solution to the original dual problem with the same objective value. Therefore, if we denote by v_P and v_D the optimal value of the original primal and dual problem respectively, then $v^\epsilon \geq v_D \geq v_P$. Let $\boldsymbol{\nu}^\epsilon$ be an optimal primal solution of the primal problem above. Define

$$\tilde{\boldsymbol{\mu}}^\epsilon := \boldsymbol{\nu}(\Xi^\epsilon)\boldsymbol{\mu}^\epsilon + \mathbb{1}_{\Xi\setminus\Xi^\epsilon}\boldsymbol{\nu}.$$

Then it holds that

$$\pi_\#^k \tilde{\boldsymbol{\mu}}^\epsilon = \boldsymbol{\nu}(\Xi^\epsilon)\pi_\#^k\boldsymbol{\mu}^\epsilon + \mathbb{1}_{\Xi\setminus\Xi^\epsilon}\pi_\#^k\boldsymbol{\nu} = \mathbb{1}_{\Xi^\epsilon}F_k + \mathbb{1}_{\Xi\setminus\Xi^\epsilon}F_k = F_k.$$

Moreover, we have that

$$W_p(\tilde{\boldsymbol{\mu}}^\epsilon, \boldsymbol{\nu}) \leq W_p(\boldsymbol{\nu}(\Xi^\epsilon)\boldsymbol{\mu}^\epsilon, \mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}) = \boldsymbol{\nu}(\Xi^\epsilon) W_p(\boldsymbol{\mu}^\epsilon, \mathbb{1}_{\Xi^\epsilon}\boldsymbol{\nu}^\epsilon) \leq \boldsymbol{\nu}(\Xi^\epsilon)x \leq x,$$

Hence $\tilde{\boldsymbol{\mu}}^\epsilon$ is feasible to the original primal problem. In addition,

$$\int_\Xi \ell d\tilde{\boldsymbol{\mu}}^\epsilon = \boldsymbol{\nu}(\Xi^\epsilon) \int_{\Xi^\epsilon} \ell d\boldsymbol{\mu}^\epsilon + \int_{\Xi\setminus\Xi^\epsilon} \ell d\boldsymbol{\nu} = \ \boldsymbol{\nu}(\Xi^\epsilon) v^\epsilon + \int_{\Xi\setminus\Xi^\epsilon} \ell d\boldsymbol{\nu}.$$

Letting $\epsilon \to 0$, we obtain that $v_P \geq v_D$. Therefore, up to a subsequence, $\tilde{\boldsymbol{\mu}}^\epsilon$ converges to $\tilde{\boldsymbol{\mu}}$, and the analysis above shows that $\tilde{\boldsymbol{\mu}}$ is primal optimal and $v_P = v_D$. □

We finally prove the measurability of the integrand involved in the dual problem. Denote by $(\Xi, \mathscr{B}_{\boldsymbol{\nu}}(\Xi), \boldsymbol{\nu})$ the completion of measure space $(\Xi, \mathscr{B}(\Xi), \boldsymbol{\nu})$ (see, e.g., Lemma 1.25 in [147]). A function $f : \mathbb{R}^m \times \Xi \to \bar{\mathbb{R}}$ is called a *normal integrand*, if the associated epigraphical multifunction $\zeta \mapsto \operatorname{epi} f(\cdot, \zeta)$ is closed valued and measurable.

Lemma C.1. *Let $f_k \in B(\Xi_k)$. The function $\Phi : \mathbb{R} \times \Xi \to \mathbb{R}$ defined by*

$$\Phi(\lambda, \zeta) := \sup_{\xi \in \Xi} \big[\ell(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta)\big]$$

is a normal integrand with respect to $\mathscr{B}(\mathbb{R}) \otimes \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$.

Proof of Lemma C.1. Define a function $g : \Xi \times \mathbb{R} \times \mathbb{R} \times \Xi \to \bar{\mathbb{R}}$ by

$$g(\xi, \lambda, \zeta) = \ell(\xi) - \sum_k f_k(\xi_k) - \lambda \mathsf{d}_F^p(\xi, \zeta).$$

Then for every $\zeta \in \Xi$, $-g(\cdot, \cdot, \cdot, \zeta)$ is lower semi-continuous, thus g is $\mathscr{B}(\Xi) \otimes \mathscr{B}(\mathbb{R}) \otimes \mathscr{B}_{\boldsymbol{\nu}}(\Xi)$-measurable. Hence by joint measurability criterion (see, e.g., Corollary 14.34 in [143]), g is a normal integrand, thereby the function Φ is also a normal integrand (Theorem 7.38 in [134]). □

REFERENCES

[1] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2014, vol. 16.

[2] H. Scarf, K. J. Arrow, and S. Karlin, "A min-max solution of an inventory problem," *Studies in the Mathematical Theory of Inventory and Production*, vol. 10, pp. 201–209, 1958.

[3] G. Gallego and I. Moon, "The distribution free newsboy problem: Review and extensions," *Journal of the Operational Research Society*, pp. 825–834, 1993.

[4] J. O. Berger, "The robust bayesian viewpoint," *Studies in Bayesian Econometrics and Statistics : In Honor of Leonard J. Savage., Edited by Joseph B. Kadane*, vol. 4, no. 2, pp. 63–124, 1984.

[5] J. Žáčková, "On minimax solutions of stochastic linear programming problems," *Časopis pro pěstování matematiky*, vol. 91, no. 4, pp. 423–430, 1966.

[6] J. Dupačová, "The minimax approach to stochastic programming and an illustrative application," *Stochastics: An International Journal of Probability and Stochastic Processes*, vol. 20, no. 1, pp. 73–88, 1987.

[7] A. Shapiro and A. J. Kleywegt, "Minimax analysis of stochastic problems," *Optimization Methods and Software*, vol. 17, no. 3, pp. 523–542, 2002.

[8] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[9] I. Popescu, "Robust mean-covariance solutions for stochastic optimization," *Operations Research*, vol. 55, no. 1, pp. 98–112, 2007.

[10] S. Zymler, D. Kuhn, and B. Rustem, "Distributionally robust joint chance constraints with second-order moment information," *Mathematical Programming*, vol. 137, no. 1-2, pp. 167–198, 2013.

[11] Z. Wang, P. W. Glynn, and Y. Ye, "Likelihood robust optimization for data-driven problems," *Computational Management Science*, pp. 1–21, 2015.

[12] J. Goh and M. Sim, "Distributionally robust optimization and its tractable approximations," *Operations Research*, vol. 58, no. 4-part-1, pp. 902–917, 2010.

[13] W. Wiesemann, D. Kuhn, and M. Sim, "Distributionally robust convex optimization," *Operations Research*, vol. 62, no. 6, pp. 1358–1376, 2014.

[14] Z. Chen, M. Sim, and H. Xu, "Distributionally robust optimization with infinitely constrained ambiguity sets," *Available on Optimization Online*, 2017.

[15] L. El Ghaoui, M. Oks, and F. Oustry, "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach," *Operations Research*, vol. 51, no. 4, pp. 543–556, 2003.

[16] G. C. Calafiore and L. El Ghaoui, "On distributionally robust chance-constrained linear programs," *Journal of Optimization Theory and Applications*, vol. 130, no. 1, pp. 1–22, 2006.

[17] G. Bayraksan and D. K. Love, "Data-driven stochastic programming using phi-divergences," *Tutorials in Operations Research*, 2015.

[18] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen, "Robust solutions of optimization problems affected by uncertain probabilities," *Management Science*, vol. 59, no. 2, pp. 341–357, 2013.

[19] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, pp. 1–37, 2015.

[20] H. Sun and H. Xu, "Convergence analysis for distributionally robust optimization and equilibrium problems," *Mathematics of Operations Research*, 2015.

[21] E. Erdoğan and G. Iyengar, "Ambiguous chance-constrained problems and robust optimization," *Mathematical Programming*, vol. 107, no. 1-2, pp. 37–61, 2006.

[22] D. Wozabal, "A framework for optimization under ambiguity," *Annals of Operations Research*, vol. 193, no. 1, pp. 21–47, 2012.

[23] ——, "Robustifying convex risk measures for linear portfolios: A nonparametric approach," *Operations Research*, vol. 62, no. 6, pp. 1302–1315, 2014.

[24] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations," *arXiv preprint arXiv:1505.05116*, 2015.

[25] C. Zhao and Y. Guan, "Data-driven risk-averse stochastic optimization with Wasserstein metric," *Available on Optimization Online*, 2015.

[26] L. V. Kantorovich, "Mathematical methods of organizing and planning production," *Management Science*, vol. 6, no. 4, pp. 366–422, 1960.

[27] L. V. Kantorovich, "On the translocation of masses," in *Dokl. Akad. Nauk SSSR*, vol. 37, 1942, pp. 199–201.

[28] G. C. Pflug and A. Pichler, *Multistage stochastic optimization*. Springer, 2014.

[29] A. Shapiro, "On duality theory of conic linear problems," in *Semi-infinite programming*, Springer, 2001, pp. 135–165.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[31] A. N. Tikhonov, V. Y. Arsenin, and F. John, *Solutions of ill-posed problems*. Winston Washington, DC, 1977, vol. 14.

[32] C. M. Bishop, "Training with noise is equivalent to tikhonov regularization," *Neural computation*, vol. 7, no. 1, pp. 108–116, 1995.

[33] S. Wager, S. Wang, and P. S. Liang, "Dropout training as adaptive regularization," in *Advances in neural information processing systems*, 2013, pp. 351–359.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[36] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[37] H. Xu, C. Caramanis, and S. Mannor, "Robust regression and lasso," in *Advances in Neural Information Processing Systems*, 2009, pp. 1801–1808.

[38] D. Bertsimas and M. S. Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 2017.

[39] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1485–1510, 2009.

[40] P. M. Esfahani and D. Kuhn, "Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations," *Mathematical Programming*, 2017.

[41] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn, "Distributionally robust logistic regression," in *Advances in Neural Information Processing Systems*, 2015.

[42] J. Blanchet, Y. Kang, and K. Murthy, "Robust wasserstein profile inference and applications to machine learning," *arXiv preprint arXiv:1610.05627*, 2016.

[43] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *arXiv preprint arXiv:1710.10016*, 2017.

[44] J.-y. Gotoh, M. J. Kim, and A. Lim, "Robust empirical optimization is almost the same as mean-variance optimization," 2015.

[45] H. Lam, "Robust sensitivity analysis for stochastic systems," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1248–1275, 2016.

[46] P. Hall and N. Neumeyer, "Estimating a bivariate density when there are extra data on one or both components," *Biometrika*, pp. 439–450, 2006.

[47] R. B. Nelsen, *An introduction to copulas*. Springer Science & Business Media, 2013, vol. 139.

[48] H. Joe, *Dependence modeling with copulas*. CRC Press, 2014.

[49] M Sklar, *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[50] W. Hoeffding, *Massstabinvariante korrelationstheorie*. Teubner, 1940, (Translated in: The CoHected Works of Wassily Hoeffding, N. I. Fisher and P. K. Sen (eds.), Springer Verlag, New York 1994).

[51] M. Fréchet, "Sur les tableaux dont les marges et des bornes sont données," *Revue de l'Institut international de statistique*, pp. 10–32, 1960.

[52] K. Natarajan, M. Song, and C.-P. Teo, "Persistency model and its applications in choice modeling," *Management Science*, vol. 55, no. 3, pp. 453–469, 2009.

[53] S. Agrawal, Y. Ding, A. Saberi, and Y. Ye, "Price of correlations in stochastic optimization," *Operations Research*, vol. 60, no. 1, pp. 150–162, 2012.

[54] X. V. Doan and K. Natarajan, "On the complexity of nonoverlapping multivariate marginal bounds for probabilistic combinatorial optimization problems," *Operations research*, vol. 60, no. 1, pp. 138–149, 2012.

[55] H. Joe, *Multivariate models and multivariate dependence concepts*. CRC Press, 1997.

[56] V. Benes and J. Stepán, *Distributions with given marginals and moment problems*. Springer Science & Business Media, 2012.

[57] F. Schmid, R. Schmidt, T. Blumentritt, S. Gaißer, and M. Ruppert, "Copula-based measures of multivariate association," in *Copula theory and its applications*, Springer, 2010, pp. 209–236.

[58] S. Dey, S. Juneja, and K. R. Murthy, "Incorporating views on marginal distributions in the calibration of risk models," *Operations Research Letters*, vol. 43, no. 1, pp. 46–51, 2015.

[59] P. Glasserman and L. Yang, "Bounding wrong-way risk in cva calculation," *Mathematical Finance*, 2016.

[60] H. Lam, "Sensitivity to serial dependency of input processes: A robust approach," *Management Science*, 2017.

[61] A. Dhara, B. Das, and K. Natarajan, "Worst-case expected shortfall with univariate and bivariate marginals," *arXiv preprint arXiv:1701.04167*, 2017.

[62] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with wasserstein distance," *arXiv preprint arXiv:1604.02199*, 2016.

[63] J. Blanchet and K. R. A. Murthy, "Quantifying distributional model risk via optimal transport," *arXiv preprint arXiv:1604.01446*, 2016.

[64] H. Owhadi and C. Scovel, "Extreme points of a ball about a measure with finite support," *arXiv preprint arXiv:1504.06745*, 2015.

[65] L. Pardo, *Statistical inference based on divergence measures*. CRC Press, 2005.

[66] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.

[67] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

[68] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[69] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 5, pp. 840–853, 2007.

[70] C. Villani, *Topics in optimal transportation*, ser. 58. American Mathematical Soc., 2003.

[71] ——, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[72] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of bounded variation and free discontinuity problems*. Oxford: Clarendon Press, 2000, vol. 254.

[73] N. Parikh and S. P. Boyd, "Proximal algorithms.," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[74] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International statistical review*, vol. 70, no. 3, pp. 419–435, 2002.

[75] A. Nemirovski, "Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[76] Y. Nesterov and A. Nemirovski, "On first-order algorithms for l 1/nuclear norm minimization," *Acta Numerica*, vol. 22, pp. 509–575, 2013.

[77] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[78] A. D. Barbour and T. C. Brown, "Stein's method and point process approximation," *Stochastic Processes and their Applications*, vol. 43, no. 1, pp. 9–31, 1992.

[79] L. H. Y. Chen and A. Xia, "Stein's method, palm theory and poisson process approximation," *Annals of Probability*, pp. 2545–2569, 2004.

[80] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*, Second. Springer, 2003, ISBN: 0-387-95541-0.

[81] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski, "Adjustable robust solutions of uncertain linear programs," *Mathematical Programming*, vol. 99, no. 2, pp. 351–376, 2004.

[82] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge Univ Press, vol. 55, 1959, pp. 299–327.

[83] F. K. Hwang and D. S. Richards, "Steiner tree problems," *Networks*, vol. 22, no. 1, pp. 55–89, 1992.

[84] L. K. Platzman and J. J. Bartholdi, "Spacefilling curves and the planar travelling salesman problem," *Journal of the ACM (JACM)*, vol. 36, no. 4, pp. 719–737, 1989.

[85] J. J. Bartholdi and L. K. Platzman, "Heuristics based on spacefilling curves for combinatorial problems in euclidean space," *Management Science*, vol. 34, no. 3, pp. 291–305, 1988.

[86] M. Haimovich and A. H. G. Rinnooy Kan, "Bounds and heuristics for capacitated routing problems," *Mathematics of Operations Research*, vol. 10, no. 4, pp. 527–542, 1985.

[87] J. M. Steele, "Subadditive euclidean functionals and nonlinear growth in geometric probability," *The Annals of Probability*, pp. 365–376, 1981.

[88] ——, *Probability theory and combinatorial optimization*. SIAM, 1997, vol. 69.

[89] J. G. Carlsson, M. Behroozi, and K. Mihic, "Wasserstein distance and the distributionally robust TSP," *submitted to Operations Reserach*, 2015.

[90] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," *arXiv preprint arXiv:1712.06050*, 2017.

[91] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[92] D. McFadden *et al.*, "Conditional logit analysis of qualitative choice behavior," *Institute of Urban and Regional Development, University of California*, 1973.

[93] D. McFadden, "Modeling the choice of residential location," *Transportation Research Record*, no. 673, 1978.

[94] D. McFadden and K. Train, "Mixed mnl models for discrete response," *Journal of applied Econometrics*, pp. 447–470, 2000.

[95] V. K. Mishra, K. Natarajan, D. Padmanabhan, C.-P. Teo, and X. Li, "On theoretical and empirical aspects of marginal distribution choice models," *Management Science*, vol. 60, no. 6, pp. 1511–1531, 2014.

[96] S. D. Ahipasaoglu, X. Li, and K. Natarajan, "A convex optimization approach for computing correlated choice probabilities with many alternatives," *Available on Optimization Online*, vol. 4034, 2013.

[97] S. P. Anderson, A. De Palma, and J.-F. Thisse, "A representative consumer theory of the logit model," *International Economic Review*, pp. 461–466, 1988.

[98] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

[99] G. Feng, X. Li, and Z. Wang, "On the relation between several discrete choice models," *Operations Research*, 2017.

[100] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with dependence structure," *arXiv preprint arXiv:1701.04200*, 2017.

[101] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[102] T. Tao, *An introduction to measure theory*. American Mathematical Soc., 2011, vol. 126.

[103] U. Shaham, Y. Yamada, and S. Negahban, "Understanding adversarial training: Increasing local stability of neural nets through robust optimization," *arXiv preprint arXiv:1511.05432*, 2015.

[104] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," *arXiv preprint arXiv:1705.09367*, 2017.

[105] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 2008, vol. 1.

[106] K. E. Train, *Discrete choice methods with simulation*. Cambridge university press, 2009.

[107] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[108] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[109] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[110] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[111] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[112] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*, Springer, 2016, pp. 597–613.

[113] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[114] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[115] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. Lanckriet, "Non-parametric estimation of integral probability metrics," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, IEEE, 2010, pp. 1428–1432.

[116] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[117] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Technical report, University of Toronto*, 2009.

[118] W. Zhang, J. Sun, and X. Tang, "Cat head detection-how to effectively exploit shape and texture features," *Computer Vision–ECCV 2008*, pp. 802–816, 2008.

[119] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[120] A. Jolicoeur-Martineau, "Deep learning with cats," *GitHub repository*, 2017.

[121] J. Dai, W. Ding, A. Kleywegt, X. Wang, and Y. Zhang, "Choice based revenue management for parallel flights," *Available on Optimization Online*, 2014.

[122] R. Gao and A. J. Kleywegt, "Data-driven robust optimization with known marginal distributions," Working paper, Tech. Rep., 2017.

[123] H. G. Kellerer, "Duality theorems for marginal problems," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 67, no. 4, pp. 399–432, 1984.

[124] S. T. Rachev, "The monge-kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[125] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume I: Theory*. Springer Science & Business Media, 1998, vol. 1.

[126] W. Gangbo and A. Swiech, "Optimal maps for the multidimensional monge-kantorovich problem," *Communications on pure and applied mathematics*, vol. 51, no. 1, pp. 23–45, 1998.

[127] J. Dhaene and M. J. Goovaerts, "On the dependency of risks in the individual life model," *Insurance: Mathematics and Economics*, vol. 19, no. 3, pp. 243–253, 1997.

[128] A. Müller, "Stop-loss order for portfolios of dependent risks," *Insurance: Mathematics and Economics*, vol. 21, no. 3, pp. 219–223, 1997.

[129] P. Deheuvels, "La fonction de dépendance empirique et ses propriétés. un test non paramétrique dindépendance," *Acad. Roy. Belg. Bull. Cl. Sci.(5)*, vol. 65, no. 6, pp. 274–292, 1979.

[130] H. Tsukahara, "Semiparametric estimation in copula models," *Canadian Journal of Statistics*, vol. 33, no. 3, pp. 357–375, 2005.

[131] R. Aldrovandi and J. G. Pereira, *An introduction to geometrical physics*. World scientific, 1995.

[132] B. Schweizer and E. F. Wolff, "On nonparametric measures of dependence for random variables," *The annals of statistics*, pp. 879–885, 1981.

[133] A. Rényi, "On measures of dependence," *Acta mathematica hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.

[134] A. Shapiro, D. Dentcheva, and A. Ruszczynski, "Lectures on stochastic programming, volume 9 of mps/siam series on optimization," *Philadelphia, PA: SIAM. Modeling and theory*, 2009.

[135] C. Zalinescu, *Convex analysis in general vector spaces*. World Scientific, 2002.

[136] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000.

[137] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, vol. 33, no. 1, pp. 3–56, 1993.

[138] K. R. Frech. (2017). 30 industry portfolios. Online; accessed April 2017.

[139] J. Fan, Y. Fan, and J. Lv, "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, vol. 147, no. 1, pp. 186–197, 2008.

[140] L. Qu and W. Yin, "Copula density estimation by total variation penalized likelihood with linear equality constraints," *Computational Statistics & Data Analysis*, vol. 56, no. 2, pp. 384–398, 2012.

[141] J.-P. Aubin and H. Frankowska, *Set-valued analysis*. Springer Science & Business Media, 2009.

[142] C. D. Aliprantis and K. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, 2006.

[143] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[144] S. Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974.

[145] F. Bolley, A. Guillin, and C. Villani, "Quantitative concentration inequalities for empirical measures on non-compact spaces," *Probability Theory and Related Fields*, vol. 137, no. 3-4, pp. 541–593, 2007.

[146] F. Bolley and C. Villani, "Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities," *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 14, no. 3, pp. 331–352, 2005.

[147] O. Kallenberg, *Foundations of modern probability*. Springer Science & Business Media, 2006.

VITA

Rui Gao was born in Luoyang, Henan Province, China in May 1992. He enrolled in the Special Class for the Gifted Young at Xi'an Jiaotong University in 2007. After a two-year preparatory program, he attended the Qian Xuesen Experimental Class at the same university for one semester and then was selected to the Honor Science Program in Mathematics. He obtained his B.Sc. in Mathematics and Applied Mathematics in 2013. Afterwards, his interest was drawn towards operations research, and he joined the Ph.D. program in the School of Industrial and Systems Engineering at Georgia Institute of Technology. With a fulfilling period and many stimulating experiences, he has completed his Ph.D. studies and now he is ready for new adventures. These will take him to The University of Texas at Austin, where he will be an assistant professor in the McCombs School of Business.